465

25 OCT 73

D-S

212

NAVAL RESEARCH LOGISTICS QUARTERLY

MONTEREY CALIF. 93940

OCT 29 1973

SEPTEMBER 1973
VOL. 20, NO. 3

OFFICE OF NAVAL RESEARCH

407-B

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORS

F. D. Rigby
Texas Tech. University

O. Morgenstern
New York University

B. J. McDonald
Office of Naval Research

S. M. Selig
Managing Editor
Office of Naval Research
Arlington, Va. 22217

ASSOCIATE EDITORS

R. Bellman, RAND Corporation
J. C. Busby, Jr., Captain, SC, USN (Retired)
W. W. Cooper, Carnegie Mellon University
J. G. Dean, Captain, SC, USN
G. Dyer, Vice Admiral, USN (Retired)
P. L. Folsom, Captain, USN (Retired)
M. A. Geisler, RAND Corporation
A. J. Hoffman, International Business Machines Corporation
H. P. Jones, Commander, SC, USN (Retired)
S. Karlin, Stanford University
H. W. Kuhn, Princeton University
J. Laderman, Office of Naval Research
R. J. Lundegard, Office of Naval Research
W. H. Marlow, The George Washington University
R. E. McShane, Vice Admiral, USN (Retired)
W. F. Millson, Captain, SC, USN
H. D. Moore, Captain, SC, USN (Retired)
M. I. Rosenberg, Captain, USN (Retired)
D. Rosenblatt, National Bureau of Standards
J. V. Rosapepe, Commander, SC, USN (Retired)
T. L. Saaty, University of Pennsylvania
E. K. Scofield, Captain, SC, USN (Retired)
M. W. Shelly, University of Kansas
J. R. Simpson, Office of Naval Research
J. S. Skoczylas, Colonel, USMC
S. R. Smith, Naval Research Laboratory
H. Solomon, The George Washington University
I. Stakgold, Northwestern University
E. D. Stanley, Jr., Rear Admiral, USN (Retired)
C. Stein, Jr., Captain, SC, USN (Retired)
R. M. Thrall, Rice University
T. C. Varley, Office of Naval Research
J. F. Tynan, Commander, SC, USN (Retired)
J. D. Wilkes, Department of Defense OASD (ISA)

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $10.00 a year in the U.S. and Canada, $12.50 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this quarterly are those of the authors and not necessarily those of the Office of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, NAVEXOS P-35

Permission has been granted to use the copyrighted material appearing in this publication.

SEQUENTIAL DETERMINATION OF INSPECTION EPOCHS FOR RELIABILITY SYSTEMS WITH GENERAL LIFETIME DISTRIBUTIONS*

S. Zacks and W. J. Fenske

Department of Mathematics and Statistics
Case Western Reserve University
Cleveland, Ohio

ABSTRACT

The problem of determining the optimal inspection epoch is studied for reliability systems in which N components operate in parallel. Lifetime distribution is arbitrary, but known. The optimization is carried with respect to two cost factors: the cost of inspecting a component and the cost of failure. The inspection epochs are determined so that the expected cost of the whole system per time unit per cycle will be minimized. The optimization process depends in the general case on the whole failure history of the system. This dependence is characterized. The cases of Weibull lifetime distributions are elaborated and illustrated numerically. The characteristics of the optimal inspection intervals are studied theoretically.

1. INTRODUCTION

In the present study we investigate the problem of determining the optimal inspection epochs of a reliability system which is comprised of N components, operating independently (in parallel) and having the same lifetime distributions. The lifetime distribution is known. An inspector visits the system at a predetermined inspection epoch and finds a certain number of components which have failed. The exact times of failure are unknown. All the components which have failed during the interval between inspections are replaced by new components. Components which have not failed are left in the system. We consider two types of cost factors: (i) The cost of inspection, which depends on the number of components in the system; and (ii) The cost of failure per unit time. This cost component measures the loss due to a failure of components. The objective is to determine an inspection policy that would be optimal with respect to the criterion of minimizing the total expected (discounted) cost for the entire future. However, since we are dealing with cases of general lifetime distributions (not necessarily exponential) the dynamic programming solution is excessively complicated, even in the truncated case (when the number of inspections should not exceed a prescribed bound). Therefore, we are considering in the present paper a sequential myopic procedure. Accordingly, after each inspection the epoch of the next inspection is determined, as a function of the whole past failure history of the system. The aim is to minimize the conditional expected cost per time unit from the present time until the next inspection epoch. In the case of exponential lifetime distributions (constant failure rates) the optimal inspection interval (time interval between inspections) does not depend on the past history of the system. As shown in the present study if the lifetime distribution is not exponential this dependence might be very strong, especially if N is not large and the lifetime distribu-

tion is of a decreasing failure rate (DFR). The dependence of the optimal inspection intervals on the observed number of failures, and on the number of components that were replaced at previous inspections and are still operating, will be explicitly characterized. We start in section 2 by formulating the model and the associated distributions. In section 3 we develop a general formula for the sequential determination of the optimal length of the inspection intervals. In section 4 we derive the corresponding formulas for lifetime distributions of the Weibull family; and illustrate the process with a numerical example. In section 5 we try to explain the complex process illustrated in the example of section 4 by further theoretical development.

There are numerous papers in the reliability literature on inspections epochs and optimal maintenance. For the general theory see chapter 4 of Barlow and Proschan [1]. Articles which are close to the present study are those of Kamins [4], Kander [5], Kander and Naor [6] and Kander and Rabinovitch [7]. The present study provides further elaboration of a chapter in the thesis of Fenske [3]. The main difference between the present study and the articles mentioned above is in the basic model. The present study is concerned with multicomponent systems while the other studies treat the whole system as one component. The study of Ehrenfeld [2] was based on a model similar to ours, but Ehrenfeld considered the problem of determining the inspection interval for the estimation of the mean time between failures in the exponential case.

2. THE MODEL AND ASSOCIATED DISTRIBUTIONS.

Consider a reliability system which consists of N, $N \geq 1$, components. These components operate independently (in parallel). Let T designate the lifetime of a component. This is a random variable having a known distribution function (*c.d.f*)$F(t)$. We assume that $F(t)$ is absolutely continuous, with a *positive* density function $f(t)$, $0 < f(t) < 0$, and $F(0) = 0$. We further assume that the expected value of T, according to $F(t)$ is finite. Let $S_0 \equiv 0$ and let $S_0 < S_1 < S_2 < \ldots < S_m < \ldots$ designate a sequence of inspection epochs. Let J_m $(m = 1, 2, \ldots)$ designate the number of components that failed during the time interval (S_{m-1}, S_m). All the J_m components are replaced at the inspection epoch S_m. The $N - J_m$ components which have not failed during (S_{m-1}, S_m) are classified into m disjoint subsets $A_0^{(m)}, A_1^{(m)}, \ldots, A_{m-1}^{(m)}$. The subset $A_j^{(m)}$ $(j=0, \ldots, m-1)$ contains all the components that were replaced at epoch S_j and did not fail throughout the time interval (S_j, S_m). Let $n_j^{(m)}$ designate the number of elements of $A_j^{(m)}$. Obviously, $A_j^{(m+1)} \subset A_j^{(m)}$ and $n_j^{(m+1)} \leq n_j^{(m)}$ for each $j = 0, 1, \ldots$ and $m = j, j+1, \ldots$. Let $n_m^{(m)} = J_m$, and $\mathbf{n}^{(m)} = (n_0^{(m)}, n_1^{(m)}, \ldots, n_m^{(m)})$ for each $m = 0, 1, \ldots$; $n_0^{(0)} \equiv n$.

If a component belongs to the subset $A_j^{(m)}$ then its conditional lifetime distribution at time t is:

$$F_j^{(m)}(t) = P\{T \leq t - S_j \mid T \geq S_m - S_j\} \tag{2.1}$$

$$= \begin{cases} 0, & \text{if } t \leq S_m \\ \dfrac{F(t - S_j) - F(S_m - S_j)}{1 - F(S_m - S_j)}, & \text{if } t > S_m. \end{cases}$$

In particular, $F_m^{(m)}(t) = F(t - S_m)$, if $t > S_m$ and zero otherwise. The conditional densities of $U = T - (S_m - S_j)$, corresponding to the life time T of a component which belongs to $A_j^{(m)}$ play an important role in our procedure. We can call U the remaining lifetime. It a component is chosen at

random at time $t = S_m + 0$ its remaining lifetime U has a conditional density function

$$h_m(u|\mathbf{S}^{(m)}, \mathbf{n}^{(m)}) = \frac{1}{N}\sum_{j=0}^{m} n_j^{(m)} \frac{f(u+S_m-S_j)}{1-F(S_m-S_j)}, \quad u \geq 0 \tag{2.2}$$

where $\mathbf{S}^{(m)} = (S_1, \ldots, S_m)$.

We notice that if T has a negative exponential distribution, i.e., $f(t)=\lambda e^{-\lambda t}$, $t \geq 0$, for any $0<\lambda<\infty$, then $h_m(u|\mathbf{S}^{(m)}, \mathbf{n}^{(m)}) = f(u)$ for *all* $m = 1, 2, \ldots$ and *all* $(\mathbf{S}^{(m)}, \mathbf{n}^{(m)})$. This is a well known property of the negative exponential distributions. Let $H_m(u|\mathbf{S}^{(m)}, \mathbf{n}^{(m)})$ designate the *c.d.f.* corresponding to (2.2).

3. SEQUENTIAL DETERMINATION OF INSPECTION EPOCHS

We will consider in the present section the problem of deriving an inspection policy which attains a certain economic objective. We assume therefore that the cost of inspecting the system is $\$C_0$ per inspection, and on the other hand, if an element fails then the cost associated with its failure is $\$C_f$ per time unit. The inspection policy adopted here is the following. Given the history connected with the past m inspection intervals, i.e., $(\mathbf{S}^{(m)}, \mathbf{n}^{(m)})$ determine the $(m+1)$ st inspection epoch so that the average expected cost per time unit of inspection and of failure, over the $(m+1)$ st inspection interval will be minimized. We remark in this connection that this policy is in essence a myopic policy, which minimizes the expected time average-costs for each inspection interval individually. A Dynamic Programming determination of the inspection epochs could attain a more global optimization. However, attempts at Dynamic Programming solutions lead to complicated sets of recursive functional equations. The solution of these equations is generally very tedious. As will be shown later the suggested myopic procedure is globally optimal if the lifetime distribution is exponential. In other cases of interest, like the Weibull distributions the myopic procedure does not coincide with the global Dynamic Programming solution. A study of the relative efficiency of the myopic procedure is still under way.

Let Δ designate the length of the $(m+1)$ st inspection interval. That is, $\Delta = S_{m+1} - S_m$. Given $(\mathbf{S}^{(m)}, \mathbf{n}^{(m)})$ the conditional expected average cost per time unit, under Δ, is

$$R_m(\Delta; \mathbf{S}^{(m)}, \mathbf{n}^{(m)}) = \frac{C_0}{\Delta} + \frac{C_f}{\Delta}\sum_{j=0}^{m} n_j^{(m)} \int_0^{\Delta} (\Delta - u) \frac{f(u+S_m-S_j)}{1-F(S_m-S_j)}\, du. \tag{3.1}$$

Or in terms of the conditional distribution of the remaining lifetime U we can express (3.1) in the form,

$$R_m(\Delta; \mathbf{S}^{(m)}, \mathbf{n}^{(m)}) = \frac{C_0}{\Delta} + NC_f H_m(\Delta|\mathbf{S}^{(m)}, \mathbf{n}^{(m)}) - \frac{C_f}{\Delta} N \int_0^{\Delta} u h_m(u|\mathbf{s}^{(m)}, \mathbf{n}^{(m)})\, du. \tag{3.2}$$

The optimal $(m+1)$ st inspection epoch is defined as $S_{m+1} = S_m + \Delta^0$, where Δ^0 is a positive real value, Δ, for which the infimum of (3.2) is attained.

Let

$$\mu_m = \int_0^{\infty} u h_m(u|\mathbf{S}^{(m)}, \mathbf{n}^{(m)})\, du, \tag{3.3}$$

be the expected remaining life, given $(\mathbf{S}^{(m)}, \mathbf{n}^{(m)})$. According to the assumption of the previous section, $\mu_m < \infty$. Differentiating $R_m(\Delta; \mathbf{S}^{(m)}, \mathbf{n}^{(m)})$ with respect to Δ we obtain that if $\mu_m \leq C_0/NC_f$ then $\Delta^0 \equiv \infty$. This is a case in which no more inspections are warranted. On the other hand, if $\mu_m > C_0/NC_f$, there exists a unique solution, Δ^0, to the equation:

$$\int_0^{\Delta} u h_m(u|\mathbf{S}^{(m)}, \mathbf{n}^{(m)})du = C_0/NC_f. \tag{3.4}$$

We realize from (3.4) that S_{m+1} is a function of the statistic $(\mathbf{S}^{(m)}, \mathbf{n}^{(m)})$ of the system.

As we have already mentioned in cases of exponential lifetime distributions the optimal length of the inspection intervals is the same for *all* $m = 1, 2, \ldots$. If $\theta = \lambda^{-1}$ is the mean time between failures (MTBF) in the exponential case then $\mu_m = \theta$ for all m, and the condition for a finite Δ^0 is that $C_0 < NC_f\theta$; i.e., the cost of inspecting an element is smaller than the expected cost of failure of an element. If this condition is satisfied then by letting $\gamma = C_0/NC_f\theta$, it is easy to show that

$$\Delta^0 = \frac{\theta}{2}\chi^2_{\gamma}[4], \tag{3.5}$$

where $\chi^2_{\gamma}[4]$ designate the γ-fractile of a chi-square distribution with 4 degrees of freedom.

4. OPTIMAL INSPECTION EPOCHS FOR WEIBULL DISTRIBUTIONS

Suppose that the lifetime of an element, T, follows a Weibull distribution, with a density function

$$f(t; \theta, \alpha) = \begin{cases} 0, & \text{if } t \leq 0 \\ \dfrac{\alpha}{\theta} t^{\alpha-1} \exp\{-t^{\alpha}/\theta\}, & \text{if } t > 0, \end{cases} \tag{4.1}$$

where α and θ are positive real parameters. We notice that if $0 < \alpha < 1$ then the distribution has a decreasing failure rate (DFR), and if $1 < \alpha < \infty$ its failure rate is increasing (IFR). When $\alpha = 1$ the distribution is exponential. Given $(\mathbf{S}^{(m)}, \mathbf{n}^{(m)})$ the density function of the remaining life U assumes the special form

(4.2)

$$h_m^{(\theta,\alpha)}(u \mid \mathbf{S}^{(m)}, \mathbf{n}^{(m)}) = \frac{1}{N}\sum_{j=0}^{m} n_j^{(m)} \exp\{(S_m - S_j)^{\alpha}/\theta\} \cdot \frac{\alpha}{\theta}(u + S_m - S_j)^{\alpha-1} \exp\{-(u + S_m - S_j)^{\alpha}/\theta\},$$

for $0 \leq u \leq \infty$. When $m = 0$ (4.2) reduces to (4.1). Following the procedure given in the previous section we realize that $S_1 < \infty$ if, and only if,

$$C_o/N < C_f\,\theta^{1/\alpha}\,\Gamma\left(\frac{1}{\alpha} + 1\right) \tag{4.3}$$

$\theta^{1/\alpha}\,\Gamma(\frac{1}{\alpha} + 1)$ is the expected lifetime. If (4.3) is satisfied then the optimal value of S_1 is

$$S_1 = \left\{ \frac{\theta}{2} G^{-1}\left(\gamma \,\middle|\, \frac{1}{2}, \frac{1}{\alpha}+1\right)\right\}^{1/\alpha}, \tag{4.4}$$

where $G^{-1}(\gamma \mid p, \nu)$, is the γ-fractile of the Gamma distribution $G(p, \nu)$, with scale parameter p, and where $\gamma = C_0/\left(NC_f \theta^{1/\alpha}\Gamma\left(\frac{1}{\alpha}+1\right)\right)$. We notice that if $2/\alpha$ is a positive integer j then

$$S_1 = \left\{ \frac{\theta}{2} \chi^2_\gamma [2+j]\right\}^{j/2}. \tag{4.5}$$

We determine now a general expression for the left-hand side of (3.4). According to (4.2),

$$\int_0^\Delta u.h_m^{(\theta,\alpha)}(u|\mathbf{S}^{(m)}, \mathbf{n}^{(m)})\, du = \frac{1}{N}\sum_{j=0}^{m} n_j^{(m)} \exp\{(S_m - S_j)^\alpha/\theta\} \cdot \tag{4.6}$$

$$\frac{\alpha}{\theta}\int_0^\Delta u(u+S_m-S_j)^{\alpha-1} \exp\{-(u+S_m-S_j)^\alpha/\theta\}\, du.$$

By a proper change of variable, we obtain

$$\frac{\alpha}{\theta}\int_0^\Delta u(u+S_m-S_j)^{\alpha-1} \exp\{-(u+S_m-S_j\}^\alpha/\theta\}\, du \tag{4.7}$$

$$= \int_{(S_m-S_j)^\alpha/\theta}^{(S_m-S_j+\Delta)^\alpha/\theta} [\theta^{1/\alpha} w^{1/\alpha} - (S_m-S_j)] \exp\{-w\}\, dw = \theta^{1/\alpha}\Gamma\left(\frac{1}{\alpha}+1\right)$$

$$\left[G\left(\frac{(S_m-S_j+\Delta)^\alpha}{\theta}; 1, \frac{1}{\alpha}+1\right) - G\left(\frac{(S_m-S_j)^\alpha}{\theta}; 1, \frac{1}{\alpha}+1\right)\right]$$

$$-(S_m-S_j)\left[\exp\left\{-\frac{(S_m-S_j)^\alpha}{\theta}\right\} - \exp\left\{-\frac{(S_m-S_j+\Delta)^\alpha}{\theta}\right\}\right].$$

Substituting (4.7) into (4.6), we obtain that $S_{m+1} = S_m + \Delta$, where Δ is the root of the equation:

$$\frac{1}{N}\sum_{j=0}^{m} n_j^{(m)} \exp\{(S_m-S_j)^\alpha/\theta\}\left[G\left(\frac{S_m-S_j-\Delta)^\alpha}{\theta}; 1, \frac{1}{\alpha}+1\right)\right. \tag{4.8}$$

$$\left. - G\left(\frac{(S_m-S_j)^\alpha}{\theta}; 1, \frac{1}{\alpha}+1\right)\right] = \gamma + \frac{1}{\theta^{1/\alpha}\Gamma\left(\frac{1}{\alpha}\right)}\frac{1}{N}\sum_{j=0}^{m} n_j^{(m)}(S_m-S_j)$$

$$\left[1-\exp\left\{-\frac{1}{\theta}[(S_m-S_j+\Delta)^\alpha - (S_m-S_j)^\alpha]\right\}\right]$$

γ is as before, and $G(x; p, \nu)$ is the c.d.f. of $G(p, \nu)$ at x.

We notice that for $m=0$ the solution of (4.8) is reduced to the one given by (4.4). In figure 1 we illustrate the solution of (4.8) for three Weibull distributions, where the $n_j^{(m)}$ sequences were generated by Monte Carlo simulation. The cases under consideration have the following parameters: $C_f=\$10.$, $C_o=\$200 \cdot N$, $\theta=$[hr] 100, and $\alpha=3/4$, 1, and 5/4. The case of $\alpha=1$ corresponds to the exponential distribution with mean $\theta=100$. According to (3.5) the optimal inspection interval for $\alpha=1$ is of length [hr] 50 χ^2_γ [4] where $\gamma=C_0/N \cdot C_f\theta=0.2$. One can find in any statistical tables that $\chi^2_{.2}$ [4] $=1.65$. Hence, the optimal interval between inspections in the exponential case is of length 82.5 hours. The case of $\alpha=5/4$ represents an IFR distribution. We see in Figure 1 that the optimal inspection intervals are of length which vary very little around 59 hours. It is interesting to notice that in the present case of an IFR distribution the optimal inspection intervals do not depend strongly on the number of components, N, in the system. This is not the case when the Weibull distribution is a DFR ($\alpha=3/4$). As illustrated in Figure 1 the optimal intervals for DFR distributions, as obtained from (4.8), are sensitive to N. When $N=10$ there are considerable fluctuations of the solution of (4.8). When $N=100$ these fluctuations diminish. The general trend of growth in the length of the inspection intervals is, however, the same. An explanation of this phenomenon will be provided in the next section. Finally we remark that the

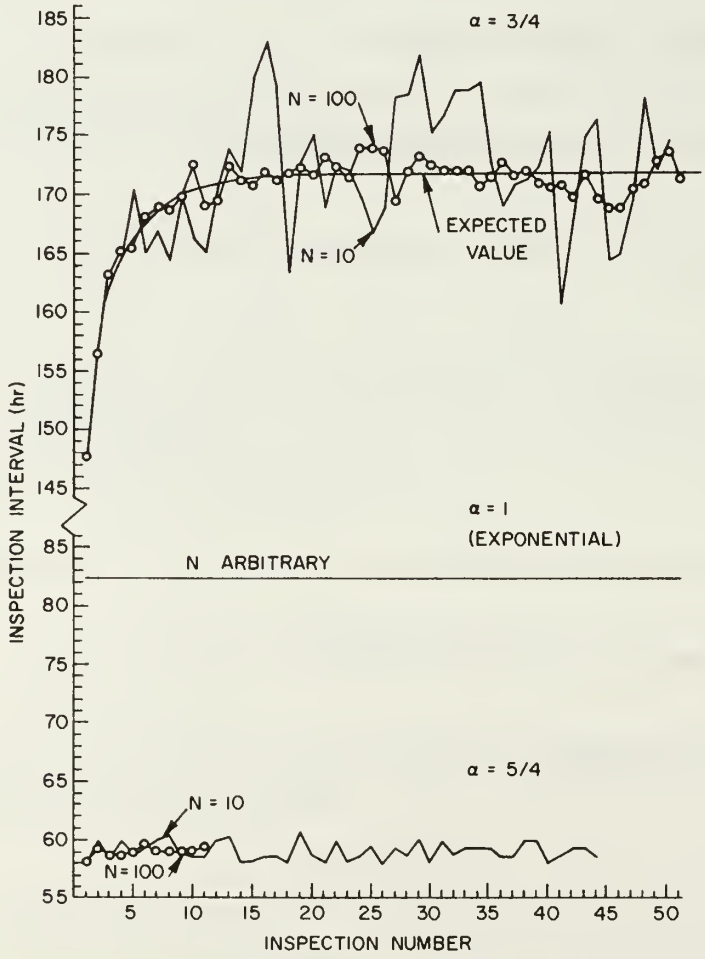

FIGURE 1. Optimum inspection intervals for Weibull Distribution with $C_f=\$10$, $C_o=\$200$N, and $\theta=100$ (hr)

numerical solution of Equation (4.8) in the case discussed here has been attained following the Newton-Raphson iterative corrections to an initial solution. For further details see Fenske [3].

5. FURTHER THEORETICAL CHARACTERIZATION OF THE SOLUTION

A characterization of the solution obtained from (3.4) is not a simple matter, since the inspection epochs $S_2, S_3, \ldots$ are random variables depending on the random vectors $\mathbf{n}^{(m)}$ in quite a complicated manner. We remark that the sequences $\{n_j^{(m)}: m=j, j+1, \ldots\}$ are supermartingales, for each $j=0, 1, 2, \ldots$; and $\lim_{m\to\infty} n_j^{(m)}$ exists as shown. Furthermore, if $S_{m-1}-S_m \geq \Delta_0$ for every m then $\lim_{m\to\infty} n_j^{(m)}=0$ for each j. This property holds for any life time distribution F. In order to obtain certain theoretical approximations to the distributions of the roots of (3.4) we consider a modified problem, in which for each $m=0, 1, \ldots$ the random variables $(n_j^{(m)}, j=0, \ldots, m)$ are replaced by some fixed non-negative values. More specifically consider the distribution of the random variable.

$$W_m=\frac{1}{N}\sum_{j=0}^{m} n_j^{(m)}[1-F(S_m-S_j)]^{-1}\int_0^{\Delta} uf(u+S_m-S_j)du, \tag{5.1}$$

in which the inspection epochs are predetermined fixed values. W_m is the left-hand side of Equation (3.4). Whenever $S_1, S_2, \ldots$ are fixed inspection epochs the vector $(n_0^{(m)}, n_1^{(m)}, \ldots, n_n^{(m)})$ has for each $m=1, 2, \ldots$ a multinomial distribution, with parameters N and $(\theta_j^{(m)}; j=0, \ldots, m)$, where $\theta_j^{(m)}$ is the probability that an element belongs to $A_j^{(m)}$, $\sum_{j=0}^{m} \theta_j^{(m)}=1$. The probabilities $\theta_j^{(m)}$ can be determined recursively according to the following formulae:

$$\theta_0^{(m)}=1-F(S_m) \tag{5.2}$$

and

$$\theta_j^{(m)}=\left(1-\sum_{i=0}^{j-1}\theta_i^{(j)}\right)[1-F(S_m-S_j)], j=1, \ldots, m.$$

It follows that for any fixed sequence of inspection epochs and for each $m=0, 1, \ldots$

$$E\{n_j^{(m)}\}=N\theta_j^{(m)}$$

$$j=0, 1, \ldots, m \tag{5.3}$$

$$\text{Var}\{n_j^{(m)}\}=N\theta_j^{(m)}(1-\theta_j^{(m)})$$

and

$$\text{cov}(n_j^{(m)}, n_k^{(m)})=-N\theta_j^{(m)}\theta_k^{(m)}, \text{all } 0\leq j<k\leq m. \tag{5.4}$$

From (5.2) and (5.3) we conclude that if the length of each inspection interval is not smaller than Δ_0 then for any distribution F, $\lim_{m\to\infty} n_j^{(m)}=0$ for each j.

The variable W_m is a linear combination of multinomial random variables. Its expectation is

$$\omega_m = E\{W_m\} = D_o^{(m)} + \sum_{j=1}^{m} \left[1 - \sum_{i=0}^{j-1} \theta_i^{(j)}\right] D_j^{(m)}, \tag{5.5}$$

where

$$D_j^{(m)} = \int_0^{\Delta} uf(u + S_m - S_j)\, du. \tag{5.6}$$

The variance of W_m is

$$\text{Var}\{W_m\} = \frac{1}{N}\left\{\sum_{j=0}^{m} \frac{\theta_j^{(m)}(D_j^{(m)})^2}{[1-F(S_m-S_j)]^2} - \left(\sum_{j=0}^{m} \frac{\theta_j^{(m)} D_j^{(m)}}{1-F(S_m-S_j)}\right)^2\right\}. \tag{5.7}$$

We have shown that for any sequence of inspection epochs $\text{Var}\{W_m\} = 0(N^{-1})$ as $N \to \infty$. This explains why the fluctuations of the roots of (4.8) are relatively large when $N = 10$ and small when $N = 100$. We consider now a particular sequence of inspection epochs which consists of values of S_m obtained by the repeated solution (for each m) of the equation $\omega_m = \gamma$, i.e.,

$$\int_0^{\Delta} uf(u)du + \sum_{j=1}^{m} \left[1 - \sum_{i=0}^{j-1} \theta_i^{(j)}\right] \int_0^{\Delta} uf(u + S_m - S_j)du = \gamma. \tag{5.8}$$

S_1 is the root Δ of $\int_0^{\Delta} uf(u)du = \gamma$, and for each $m = 1, 2, \ldots$, the $(m+1)$st inspection epoch is given by $S_{m+1} = S_m + \Delta$. The sequence of fixed inspection epochs determined by this procedure corresponds to the expected values of $n^{(m)}$ and we therefore label this procedure as the *Procedure Of Averages*. In Table 1 we provide the inspection intervals determined by the Procedure of Averages, and the corresponding multinomial probabilities $\theta_j^{(m)}$ $(j = , \ldots, m,)$, for the two cases represented in Figure 1. The graph of the corresponding inspection intervals for the case of $\alpha = 3/4$ (DFR) is also plotted in Figure 1. As is demonstrated in Table 1, in the IFR case $(\alpha = 5/4)$ the significant contribution to the solution is expected to be that of $n_m^{(m)}$ and $n_{m-1}^{(m)}$, or of their corresponding expected values. Furthermore, the optimal length of the inspection intervals varies very little with the number of inspections, m, and its expectation reaches in the present example a stable situation after two inspections. This is not the case, however, in the DFR distribution $(\alpha = 3/4)$. The probabilities $\theta_j^{(m)}$ approach zero, as m grows, very slowly. This is reflected in a steady increase in the length of the inspection intervals as m grows, and a stable situation is reached in the present example only after 10 inspections.

To insure that the inspection intervals discussed in sections 3 and 4 will have similar properties to those determined by procedures of fixed inspection epochs we could consider the following adjustment. First, determine for each $m = 1, 2, \ldots$ two fixed sequences of inspection epochs which will constitute upper and lower (confidence) limits for the solution of (3.4) (or (4.8)). This can be done by utilizing formulae (5.5) and (5.7). The lower confidence limits could be obtained by repeated solution (for the root Δ) of the equation

TABLE 1. *Values of optimal inspection intervals* Δ [*hr*] *and multimonial probabilities under the Procedure Of Averages for Weibull distributions with* $\theta = 100$ [*hr*] *and cost components* $C_o =$ *\$200*N, $C_f =$ *\$10*

Case I: $\alpha = 5/4$ (IFR)

m	opt. Δ	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$	$j=10$
1	58.0	0.2017	0.7983									
2	59.3	0.0211	0.1543	0.8246								
3	59.1	0.0016	0.0161	0.1600	0.8223							
4	59.1	0.0010	0.0013	0.0167	0.1595	0.8225						
5	59.1	0	0.0001	0.0013	0.0166	0.1595	0.8225					
6	59.1	0	0	0.0001	0.0013	0.0166	0.1595	0.8225				
Case II: $\alpha = 3/4$ (DFR)												
1	147.6	0.6548	0.3452									
2	156.8	0.4825	0.2216	0.2959								
3	161.8	0.3666	0.1624	0.1880	0.2830							
4	164.8	0.2839	0.1231	0.1372	0.1787	0.2771						
5	166.6	0.2229	0.0953	0.1039	0.1302	0.1742	0.2735					
6	167.9	0.1769	0.0748	0.0803	0.0984	0.1267	0.1716	0.2713				
7	168.8	0.1416	0.0594	0.0630	0.0761	0.0957	0.1246	0.1698	0.2698			
8	169.4	0.1142	0.0475	0.0500	0.0600	0.0739	0.0941	0.1233	0.1686	0.2686		
9	169.9	0.0927	0.0383	0.0401	0.0473	0.0580	0.0726	0.0930	0.1223	0.1678	0.2678	
10	170.3	0.0756	0.0311	0.0323	0.0379	0.0460	0.0569	0.0718	0.0924	0.1216	0.1671	0.2673

(5.9) $$\omega_m + 3.\,[\mathrm{Var}\,\{W_m\}]^{1/2} = \gamma; \qquad m = 1,2,\ldots.$$

The upper limit can be obtained by solving the equation

(5.10) $$\omega_m - 3.\,[\mathrm{Var}\,\{W_m\}]^{1/2} = \gamma, \qquad m = 1,2,\ldots.$$

In the second phase of computation solve Equation (3.4). If the solution lies between the roots of (5.9) and (5.10) proceed; otherwise truncate the solution to either the lower limit or to the upper limit, whichever is closer to the actual solution. Such an adjustment will guarantee that every inspection interval will be bounded by lower and upper values which are determined by fixed sequences of inspection epochs, and will therefore have general characteristics as established here.

REFERENCES

[1] Barlow, R. E. and F. Proschan, *Mathematical Theory of Reliability* (John Wiley and Sons, New York, 1967).

[2] Ehrenfeld, S. "Some Experimental Design Problems in Attribute Life Testing," J. Am. Stat. Assoc., *57*, 668–679 (1962).

[3] Fenske, W. J., "Optimal Inspection Epochs For Reliability Studies" Ph.D. Dissertation, Department of Mathematics and Statistics, Case Western University (1972).

[4] Kamins, M., "Determining Checkout Intervals for Systems Subject to Random Failures," The Rand Corporation, Paper RM–2578 (1960).

[5] Kander, Z., "Inspection Policies of Deteriorating Equipment Characterized by N Quality Levels," Technion-Israel Institute of Technology, Operation Research Monograph No. 93 (1971).

[6] Kander, Z. and P. Naor, "Optimization of Inspection Policies by Dynamic Programming, Technion-Israel Institute of Technology, Operation Research Monograph No. 61 (1970).

[7] Kander, Z. and A. Rabinovitch, Maintenance Policies When Failure Distribution of Equipment is Only Partially Known, Technion-Israel Institute of Technology, Operations Research Monograph No. 92 (1972).

AN EMPIRICAL BAYES ESTIMATOR FOR THE SCALE PARAMETER OF THE TWO-PARAMETER WEIBULL DISTRIBUTION

G. Kemble Bennett

Virginia Polytechnic Institute and State University

and

H. F. Martz

Texas Tech University

ABSTRACT

An empirical Bayes estimator is given for the scale parameter in the two-parameter Weibull distribution. The scale parameter is assumed to vary randomly throughout a sequence of experiments according to a common, but unknown, prior distribution. The shape parameter is assumed to be known, however, it may be different in each experiment. The estimator is obtained by means of a continuous approximation to the unknown prior density function. Results from Monte Carlo simulation are reported which show that the estimator has smaller mean-squared errors than the usual maximum-likelihood estimator.

INTRODUCTION

A large number of authors have considered estimation of the parameters of the Weibull distribution by the method of maximum likelihood, the method of moments, and numerous other classical techniques. Frequently, however, the parameters of the Weibull distribution are subject to random variation and an analysis which encompasses this feature is best suited. Such an analysis has been performed by Soland for the cases where the scale parameter is treated as a random variable [5] and where both the shape and scale parameters are treated as random variables [6]. These approaches exhibit a Bayesian viewpoint as adequately described by Raiffa and Schlaifer [4]. Emphasis is placed on determining conjugate prior distributions and on performing both terminal and preposterior analysis. In this paper we obtain empirical Bayes estimates for the scale parameter. This approach, like the Bayesian approach, allows for the assumption of a randomly varying scale parameter. The analysis, however, does not require any specific assumptions as to the distributional form of this parameter. Since this distribution generally remains unknown, the empirical Bayes approach can, in a large majority of cases, be successfully applied. Application would certainly be warranted, for example, in reliability life testing situations where the lifetime distribution of items subjected to routine testing is adequately described by a Weibull disbribution but where the scale parameter varied from test-to-test.

THE EMPIRICAL BAYES APPROACH

Consider the situation in which we observe a value t (which may be vector valued) from a Weibull density function given by

$$f(t|\lambda) = \lambda\beta t^{\beta-1} e^{-\lambda t^{\beta}} \tag{1}$$

and must estimate the parameter λ with small squared-error. The shape parameter, β, is assumed to be known, however, the scale parameter, λ, which determines t, is itself assumed to be a realization of an unobservable random variable. Furthermore, it is assumed that this estimation process is a routinely reoccurring situation. Therefore, as the process is repeated we obtain a sequence of realizations of independent and identically distributed random variables $t_1, t_2, \ldots, t_n$. Our problem is to determine an estimator $\tilde{\lambda}_n = \tilde{\lambda}_n(t_1, \ldots, t_n)$ *which minimizes* $E(\tilde{\lambda}_n - \lambda_n)^2$, where the expectation is taken with respect to all the random variables involved and where λ_n is the nth or current realization of λ. Since λ is itself a random variable this minimizing estimator is well known to be the Bayes estimator, the mean of the posterior distribution. This estimator can be represented by

$$E(\lambda|t) = \frac{\int \lambda f(t|\lambda) g(\lambda) d\lambda}{\int f(t|\lambda) g(\lambda) d\lambda}, \tag{2}$$

where $g(\lambda)$ is the true prior density function of λ.

Since the prior density usually remains unknown in practice, the estimator, $E(\lambda|t)$, cannot be exactly determined. It can, however, be approximated using the information, $t_1, t_2, \ldots, t_n$, obtained from previous realizations of λ. Such an estimator is commonly referred to as an empirical Bayes estimator. For a detailed discussion of the empirical Bayes approach the reader is referred to, for example, Maritz [2].

To illustrate this situation, consider a repetitive testing program in which the time-to-failure density of tested items is given by (1). During each test a sample of k failure times is observed from (1) and an estimate of λ is to be given. For example, at the first test a sample of k failure times is recorded and an estimate of λ_1 is required. At the second test an additional random sample of k failure times is obtained and an estimate of λ_2, which may be different from λ_1, is required. This situation continues until the present or nth test is completed, at which time an estimate of λ_n is to be given. Due to changing environmental conditions from test to test, imperfect testing equipment, interactions of population components, etc., the values $\lambda_1, \lambda_2, \ldots, \lambda_n$ are not likely to be equal, but to vary unpredictably and thus randomly. This variation can therefore be described by a prior density function. However, since the values of λ remain unknown specification of $g(\lambda)$ can often be risky. In the situation described here the observed experimental data, $t_1, t_2, \ldots, t_n$, are used to approximate $g(\lambda)$, thereby relieving the experimenter of the task of specifying the form of $g(\lambda)$.

DEVELOPMENT OF THE ESTIMATOR

Suppose that at each experiment j comprising a testing program a random sample from (1) is taken and a maximum likelihood estimate

$$\hat{\lambda}_k = k / \sum_{i=1}^{k} t_i^{\beta} \tag{3}$$

formed. Then the sequence of estimates

$$\hat{\lambda}_{k,1}, \hat{\lambda}_{k,2}, \ldots, \hat{\lambda}_{k,n} \tag{4}$$

provides a source of information on the past behavior of λ. Based on this sequence of linear transforma-

tion can be performed which yields a new sequence of values

$$\lambda_1^*, \lambda_2^*, \ldots, \lambda_n^*, \tag{5}$$

which when considered collectively have a mean and variance approximating those of the random parameter λ. This sequence can now be used to approximate the prior density function, $g(\lambda)$. Proper substitution of the approximation into (2) will yield an empirical Bayes estimate for λ_n, the nth or present realization of λ.

The particular density approximation chosen is described by Parzen [3]. He presents a consistent density estimator of the form

$$g_n(\lambda) = \frac{1}{nh(n)} \sum_{i=1}^{n} W\left(\frac{\lambda - \lambda_i}{h(n)}\right), \tag{6}$$

where $W(\cdot)$ is a weighting function satisfying certain boundedness and regularity conditions, and $h(n)$ a smoothing constant so chosen that $\lim_{n\to\infty} h(n) = 0$ and $\lim_{n\to\infty} nh(n) = \infty$. These restrictions are placed on $h(n)$ to assure the consistency of $g_n(\lambda)$. Parzen also lists several possible representations for $W(\cdot)$. Using these results Bennett and Martz [1] suggest the replacement of each unobservable λ_i in (6) by its corresponding transformed estimate λ_i^* and form the density approximation

$$\hat{g}_n(\lambda) = \frac{1}{nh(n)} \sum_{i=1}^{n} W\left(\frac{\lambda - \lambda_i^*}{h(n)}\right). \tag{7}$$

Subsequent substitution of (7) into (2) yields the empirical Bayes estimator

$$E_n(\lambda|t) = \frac{\int \lambda f(t|\lambda)\hat{g}_n(\lambda)d\lambda}{\int f(t|\lambda)\hat{g}_n(\lambda)d\lambda}. \tag{8}$$

The maximum likelihood estimator, $\hat{\lambda}_k$, given by (3) is well known to be both consistent and sufficient for estimating λ and can be easily shown to have the conditional density function

$$f(\hat{\lambda}_k|\lambda) = [k\lambda/\hat{\lambda}_k]^{k+1} \exp[-k\lambda/\hat{\lambda}_k]/\Gamma(k)\lambda k \tag{9}$$

with mean and variance given by

$$E(\hat{\lambda}_k|\lambda) = \lambda k/(k-1) \text{ and } \operatorname{Var}(\hat{\lambda}_k|\lambda) = (\lambda k)^2/(k-1)^2(k-2), \tag{10}$$

respectively. Since the maximum likelihood estimator is sufficient for estimating λ the Bayes estimator $E(\lambda|t)$, as defined by (2) can be conveniently written as $E(\lambda|\hat{\lambda}_k)$, and the corresponding empirical Bayes estimator becomes

$$E_n(\lambda|\hat{\lambda}_k) = \frac{\int \lambda f(\hat{\lambda}_k|\lambda)\hat{g}_n(\lambda)d\lambda}{\int f(\hat{\lambda}_k|\lambda)\hat{g}_n(\lambda)d\lambda}, \tag{11}$$

where $f(\hat{\lambda}_k|\lambda)$ and $\hat{g}_n(\lambda)$ are given by (9) and (7), respectively.

Since the actual range of λ remains unknown the region of integration in (8) and (11) must be determined empirically. This can be satisfactorily resolved by taking the region of integration to be the observed range of estimates upon which the prior density approximation is based. Thus, it is only necessary to order successively the estimates $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \lambda_n$ to obtain the region of integration. Alternatively, the positive half of the real line could be used.

Let us now consider the linear transformation of $\hat{\lambda}_k$ defined by

$$\lambda^* = C^{1/2}[\hat{\lambda}_k - E(\hat{\lambda}_k)] + E(\lambda), \tag{12}$$

where

$$C = \text{Var}\ (\lambda) / \text{Var}\ (\hat{\lambda}_k). \tag{13}$$

The mean and variance of λ^* are easily verified to be equivalent to those of the random variable λ, e.g., $E(\lambda^*) = E(\lambda)$ and Var $(\lambda^*) =$ Var (λ). If the mean and variance of λ were known then the transformation could be applied to each of the maximum likelihood estimates of sequence (4) forming sequence (5). Since the mean and variance of λ generally remain unknown, estimates of these quantities are required. Using relationships of conditional probability, we have from (10) that

$$E(\hat{\lambda}_k) = E[E(\hat{\lambda}_k \mid \lambda)] = [k/(k-1)]E(\lambda) \tag{14}$$

and that

$$\begin{aligned} \text{Var}\ (\hat{\lambda}_k) &= \text{Var}\ [E(\hat{\lambda}_k \mid \lambda)] + E[\text{Var}\ (\hat{\lambda}_k \mid \lambda)] \\ &= [k/(k-1)]^2\ \text{Var}\ (\lambda) + [k/(k-1)^2\ (k-2)]E^2(\lambda). \end{aligned} \tag{15}$$

From (14) the prior mean can be consistently estimated by

$$\widehat{E(\lambda)} = [(k-1)/k]\ \bar{\lambda}_n, \tag{16}$$

where $\bar{\lambda}_n$ denotes the sample mean $\frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}_{k,i}$. If in (15) Var $(\hat{\lambda}_k)$ is replaced by the sample variance, $S_n^2 = \sum_{i=1}^{n} (\hat{\lambda}_{k,i} - \bar{\lambda}_n)^2/n$, $E^2(\lambda)$ is replaced by the relation, $E(\lambda^2) = E^2(\lambda) + \text{Var}\ (\lambda)$, and $E(\lambda)$ is replaced by $\bar{\lambda}_n$, then upon solving for VAR (λ), we obtain

$$\widehat{\text{Var}}\ (\lambda) = [(k-2)S_n^2 - \bar{\lambda}_n^2]\ (k-1)/k^2 \tag{17}$$

as an estimate of Var (λ). Proper substitution of the above results into (12) yields

$$\lambda^* = C^{1/2}[\hat{\lambda}_k - \bar{\lambda}_n] + [(k-1)/k]\bar{\lambda}_n, \tag{18}$$

where

$$C = (k-1)/k[(k-2) - \lambda_n^2/S_n^2]. \tag{19}$$

Thus, the transformation defined by (12) is completely determined and the empirical Bayes estimator given by (11) can be formed.

MONTE CARLO SIMULATION

To ascertain the usefulness of the empirical Bayes estimator, $E_n(\lambda|\hat{\lambda}_k)$, Monte Carlo simulation was employed on a UNIVAC 1108 computer. The criterion of comparison chosen was mean-squared error, and the widely utilized maximum likelihood estimator was the measurement reference. Therefore, the ratio

$$R = \frac{\text{empirical Bayes mean-squared error}}{\text{maximum likelihood mean-squared error}} \tag{20}$$

was of interest.

In the simulation, a value of λ was randomly generated from a chosen prior density function selected from the Pearson family of distributions [7]. Then a random sample $t_1, t_2, \ldots, t_k$ of size k, corresponding to the realization λ was obtained from (1). The maximum likelihood estimate $\hat{\lambda}_k$ was then computed from (3) and its squared deviation $(\lambda - \hat{\lambda}_k)^2$ from the corresponding realization of λ was calculated. For the second experiment, a new value of λ was generated and the process repeated, obtaining $\hat{\lambda}_k$ and its squared deviation. For this experiment, $E_2(\lambda|\hat{\lambda}_k)$ and its squared deviation, $[\lambda - E_2(\lambda|\hat{\lambda}_k)]^2$, from the corresponding realization of λ were also calculated. This was repeated 20 times, and each time, $E_n(\lambda|\hat{\lambda}_k)$ was calculated using the present $\hat{\lambda}_{k,n}$ as well as all previous maximum-likelihood estimates. Five hundred repetitions of this run of 20 experiments were then made, and the averages of the squared deviations of $\hat{\lambda}_k$ and $E_n(\lambda|\hat{\lambda}_k)$ were formed as estimates of $E(\lambda - \hat{\lambda}_k)^2$ and $E[\lambda - E(\lambda|\hat{\lambda}_k)]^2$, respectively. Then the ratio R was calculated utilizing these estimated mean-squared errors. All numerical integrations were performed by means of the 11-point Gauss quadrature formula and the weighting function $W(\cdot)$ in (7) was taken to be

$$W(Y) = \left[\frac{\operatorname{Sin} Y}{Y}\right]^2,$$

where

$$Y = (\lambda - \lambda_i^*)/2h(n) \text{ and } h(n) = n^{-1/5}.$$

This procedure was repeated for all types of Pearson prior distributions and the ratio R was observed to be significantly influenced by the prior distribution only through the ratio of the conditional variance of the maximum likelihood estimator to the prior variance of λ. This value can be represented by

$$Z = \frac{k^2 E^2(\lambda)}{(k-1)^2 (k-2) \operatorname{Var}(\lambda)}, \tag{21}$$

where $E(\lambda)$, the prior mean of λ, has been substituted for λ. Since the only factors affecting the ratio R, apart from the number of experiences, are contained in (21) this quantity can be conveniently used to summarize and index a given situation.

It was generally observed in the simulation that the ratio R varied only slightly for a given value of n, providing the value of Z remained invariant from distribution to distribution. These results indicate the robustness of the smooth empirical Bayes estimator to the form of the prior distribution. Also, it was observed that as Z increases, the values of the ratio R decrease. This phenomenon is best understood by considering the summary quantity Z. If Var $(\hat{\lambda}_k \mid \lambda)$ is large as compared to Var (λ), then the maximum likelihood estimate of λ will vary widely. The empirical Bayes estimator, however, is capable of detecting this variation and can use this information to obtain better estimates of λ. Conversely if Var $(\hat{\lambda}_k \mid \lambda)$ is small as compared to Var (λ), then the maximum likelihood estimator would be expected to do quite well. In this case there is a great deal of information within an experiment, and previous experiments contribute very little information about the parameter. This improvement, however, never surpasses that achieved by the empirical Bayes estimator.

Values of R are plotted, in figure 1, as a function of n, the number of past experiences, for different values of Z ranging from 0.5 to 5.0. For ease of presentation curves have been smoothed through the actual data points.

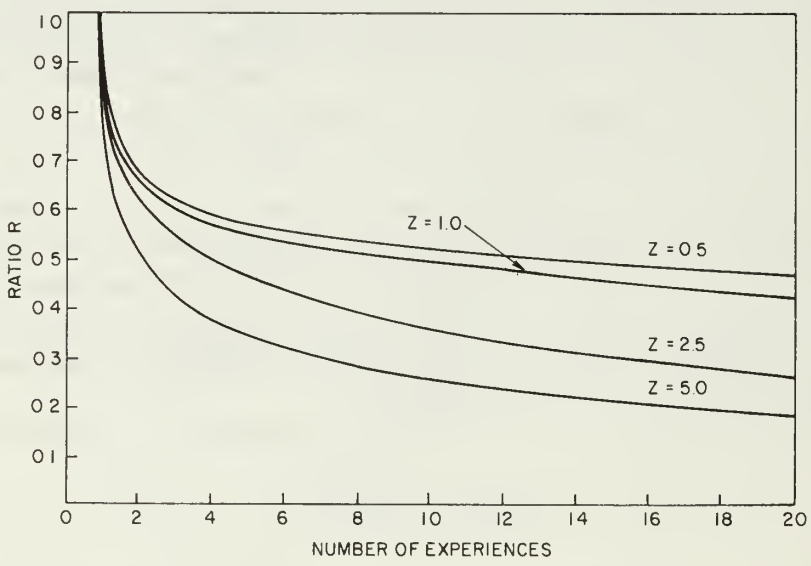

FIGURE 1. Ratio of the average squared-error of $E_n(\lambda \mid \hat{\lambda}_k)$ to the average squared-error of $\hat{\lambda}_k$ for several values of Z.

Figure 2 illustrates a typical comparison between the improvements realized from using the linear transformation defined by (18) and by not incorporating this feature into the analysis. The dotted line represents the ratio R formed with an empirical Bayes estimator whose prior density approximation is directly based on the sequence of maximum likelihood estimates (4). The solid line represents the ratio R formed with the empirical Bayes estimator as defined by (11). Here the prior density approximation is based on the transformed sequence given in (5). Both ratios illustrate the improvements over the maximum likelihood estimator achieved by both empirical Bayes estimators. Note, however, that the solid line is significantly lower than the dotted line. This result was repeatedly reproduced in the simulation for all values of Z considered.

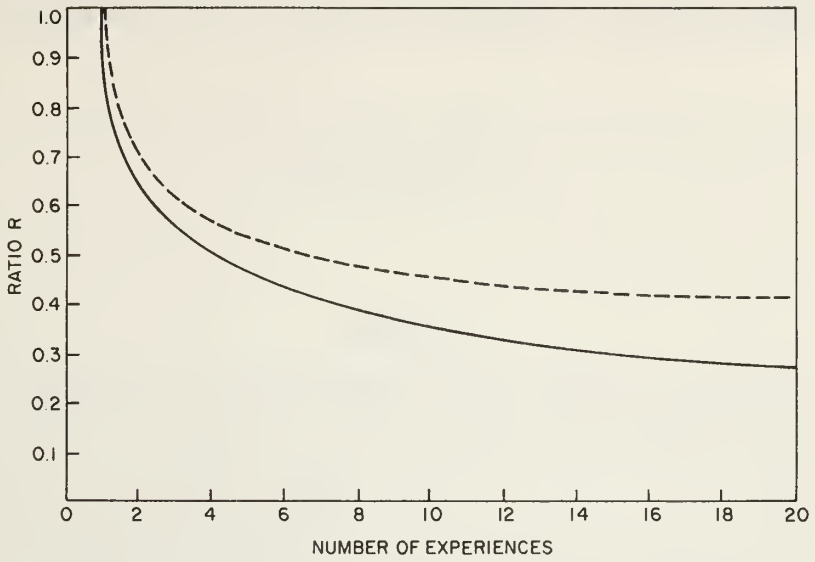

FIGURE 2. Typical comparison of the ratio R with and without the linear transformation on $\hat{\lambda}_k$; (———) with, (----------) without.

REFERENCES

[1] Bennett, G. K., and H. F. Martz, "A Continuous Empirical Bayes Smoothing Technique," Biometrika, 59, 2, 361–368, (1972).

[2] Maritz, J. S., *Empirical Bayes Methods* (Methuen and Co., Ltd., London, England, 1970).

[3] Parzen, E., "On Estimation of a Probability Density Function and Mode," Ann. Math. Statist. *33*, 1065–1076 (1962).

[4] Raiffa, H. and R. Schlaifer, *Applied Statistical Decision Theory* (Harvard Graduate School of Business Administration, 1961).

[5] Soland, R. M., "Bayesian Analysis of the Weibull Process With Unknown Scale Parameter and Its Application to Acceptance Sampling," IEEE Trans. Reliability *R–17*, 84–90 (1968).

[6] Soland, R. M., "Bayesian Analysis of the Weibull Process With Unknown Scale and Shape Parameters," IEEE Trans. Reliability *R–18*, 181–184 (1969).

[7] Thomas, D. G., "Computer Methods for Generating Pseudo-Random Numbers from Pearson Distributions and Mixtures of Pearson and Uniform Distributions," Unpublished Master of Science Thesis, Virginia Polytechnic Institute and State University (1966).

OPTIMAL ALLOCATION OF UNRELIABLE COMPONENTS FOR MAXIMIZING EXPECTED PROFIT OVER TIME

Claude G. Henin

Faculty of Management Sciences
University of Ottawa
Canada

ABSTRACT

In the present paper, we solve the following problem: Determine the optimum redundancy level to maximize the expected profit of a system bringing constant returns over a time period T; i.e., maximize the expression $P\int_0^T Rdt - C$, where P is the return of the system per unit of time, R the reliability of this system, C its cost, and T the period for which the system is supposed to work.

We present theoretical results so as to permit the application of a branch and bound algorithm to solve the problem. We also define the notion of consistency, thereby determining the distinction of two cases and the simplification of the algorithm for one of them.

1. INTRODUCTION

In [4] we described different methods for solving the following problem. A serial system made of n independent stages, has a reliability of R and a cost of C, for a mission of a certain duration. If the system functions throughout the whole mission, the resultant revenue is P dollars. The problem therefore was to maximize the expression $PR-C$, where R and C are increasing functions of the number of standby units at each stage.

In this paper, we will consider a similar but more practical problem: we will suppose that, when working, the system produces certain revenue per unit of time. This seems to be more likely to happen in real life problems. For example, let us consider the orbiting of a commercial satellite or the placement of a submarine cable, which are sources of continual revenue as long as they function properly. In both cases, the reliability of the system can be increased significantly through redundancy before the system begins operations; however the system can only be repaired with difficulty once it fails. The reliability of the system is equal to the product of the reliabilities, R_i, of each stage i. At each stage, we have m_i components; i.e., one basic unit and (m_i-1) standbys. At each instant t^* the reliability R_i is an increasing function of m_i, the number of components.

The cost of a stage is m_ic_i, where c_i is the acquisition cost of one component of type i, and the system returns a net revenue of P per unit of time, while functioning. The problem is to maximize the profit for a period T (where T can be infinite); i.e., to maximize the expression:

$$(1) \qquad P\int_0^T R(m_1, \ldots, m_n; t)dt - \sum_{i=1}^{n} c_i m_i,$$

where $R(m_1, \ldots, m_n; t) = \prod_{i=1}^{n} R_i(m_i, t)$.

*The reliability $R_i(t)$ of stage i is the probability that stage i is still working by time t. We neglect the influence of switching devices which could diminish the reliability of each stage r_i, if m_i becomes too large.

The functions $R_i(m_i, t)$ are non-increasing in t because the reliability of each stage cannot increase with time.

Another problem arises if we suppose that the returns are discounted at a rate r. In this case, the problem becomes

$$\max P \int_0^T e^{-rt} R(m_1, \ldots, m_n; t)dt - \sum_{i=1}^{n} c_i m_i. \tag{2}$$

But it is clear that (2) is identical to (1) with the factor e^{-rt} included in R; i.e., if we replace R in (1) by $R' = e^{-rt}R$. For computational reasons, it is easier to replace each $R_i (i = 1, \ldots, n)$ by $R_i' = e^{-rt/n}R_i'$, which has the same effect as replacing R by R'. With this last transformation the treatment of the two problems is identical and we shall restrict ourselves to the analysis of the first one. We shall analyze the properties of function (1) and indicate how this problem can be solved with the help of the methods described in [4].

2. CONSISTENCY

In order to analyze the function (1), we must define an important property. Let $\mathbf{m}$ denote the n-vector $m_1, \ldots, m_n$, and let us consider a subset S of indices among all the possible $\mathbf{m}$. This subset S is said to be consistent with respect to the failure law of the component and the structure of the system if the following property holds: if for some and for some $\mathbf{m}, \mathbf{m}' \epsilon S, t \leq T, R(\mathbf{m}, t) > > R(\mathbf{m}', t$, then $R(\mathbf{m}, t') \geq R(\mathbf{m}', t')$ for all $t' \leq T$. Consistency for a set S implies that the reliability orderings among all the vectors belonging to S remain constant between zero and T.

By taking T sufficiently small and by taking an upperbound N on the number of components at each stage i, it is always possible to find consistent reliabilities. Indeed, if we have n stages, we have at most N^n reliability functions. As this number is finite, it is always possible to find consistent reliabilities for $[O, T]$, where T is smaller than the smallest positive intersection point of these N^n functions.

It seems impossible to provide general necessary and sufficient conditions to insure consistency on the whole set of reliabilities. However, if the reliability function at a given stage $R_i(m, t)$ can be written as $P_i(m)g_i(t)$, then the reliabilities are consistent.

On the other hand, it is easy to find nonconsistent reliability functions. Numerical tests show that loaded standbys with exponential failure laws do not show consistent reliabilities for general values of T. As another case, consider a two stage system with components in a loaded mode and having a linear failure law; i.e., the reliability of the system is: $R = [1 - (\lambda_1 T)^{m_1}][1 - (\lambda_2 t)^{m_2}]$. We shall show that for $\lambda_1 \geq \lambda_2$, we can find two reliability curves which intersect each other. It is sufficient, for example, to take $\mathbf{m} = (m_1, m_2) = (4, 1) \mathbf{m}' = (m_1', m_2') = (2, 2)$. T must be less than $1/\lambda_1$. For $t > \lambda_2/\lambda_1^2$, $R(\mathbf{m}, t)$ is larger than $R(\mathbf{m}', t)$, but for $t < \lambda_2/\lambda_1^2$, it is smaller. Therefore, for $\lambda_2/\lambda_1^2 < T < 1/\lambda_1$, the two reliability curves intersect each other and are inconsistent. For T smaller than λ_2/λ_1^2, they do not intersect and are consistent. For larger number of stages, it becomes very difficult (computationally) to see if the reliability curves intersect each other.

Therefore, the problem exposed in this paper is more difficult than the one solved in [4]. For example, generally there is no sense in creating an undominated sequence of allocations at a given time as in [5]. However, if we compute an undominated sequence at a given time, and if the terms (i.e., the vectors $\mathbf{m}$ of this optimal sequence) are consistent for all t between 0 and T for all possible sequences, then we have the following theorem:

THEOREM I: If the set of vectors m of an undominated sequence at a given time, t, is consistent (i.e., if the undominated sequence remains the same for all t) then the optimal solution to (1) corresponds to a term of this undominated sequence.

PROOF: Suppose the contrary and that the optimal solution is given for a vector $\mathbf{m}$ not belonging to the optimal sequence and with a cost of C. Consider the two successive terms in the optimal sequence of cost C' and C'' such that $C' < C \leqslant C''$. Then by definition of the optimal sequence and consistency, we have that the reliability $R(t)$ of our solution is always smaller than the corresponding reliability $R'(t)$ of the term costing C' in the optimal sequence. Therefore, $P\int_0^T R(t)dt - C$ is smaller than $P\int_\infty^T R'(t)dt - C'$ and we arrive at a contradiction.

If the optimal sequence varies from one point in time to another, it is not true that the optimal solution to (1) always is a term in any optimal sequence. It can be a term which never appears in any optimal sequence at any time. Therefore, the use of optimal sequence is interesting only when they are identical at any time between zero and T. By extension, we shall call such an optimal sequence consistent.* As shown by our former example, it is very difficult to establish the consistency for two reliability curves even in the case of a very simple failure law. A fortiori, it is very difficult if not impossible to establish such a property for a set, such as an optimal sequence, which cannot be formally defined mathematically. Therefore, the only way of checking the consistency of an undominated sequence is to compute optimal sequences for different times and see if they are identical. If they are for a reasonable number of trials, it can be assumed that the undominated sequence is consistent. Naturally consistency on the optimal sequence is a far weaker restriction than consistency on the whole set of reliabilities.

3. PROPERTIES OF THE FUNCTION

If the standbys are in a loaded mode and if the unreliability of a component at stage i is $p_i(t)$, formula (1) becomes

$$P\int_0^T \prod_{i=1}^{n} \{1-[p_i(t)]^{m_i}\}dt - \sum_{i=1}^{n} m_i c_i = P\int_0^T R(m,t)dt - C$$

by taking $C = \sum_{i=1}^{n} m_i c_i$. (see [4].)

For such a situation, we have the following properties. Proofs from previous chapters can be used to verify these properties.

PROPERTY I: If the standby units at each stage are in a loaded mode, if $p_i(t) \geqslant p_j(t)$ for all t between 0 and T and if $c_i < c_j$, then at the optimal solution, the optimal number of components at stage i, m_i^* is larger than or equal to the optimal number of components at stage j, m_j^*.

PROOF: See the proof of Theorem I in [3].

This property reduces the number of solutions we must consider when using a branch and bound technique.

*This property implies more than consistency on the set of indices of the optimal sequence and less than consistency for all the possible indices.

PROPERTY II: If P increases, $\int_0^T Rdt$ and C are nondecreasing.

PROOF: As the proof of Theorem VI in [4].

COROLLARY: If the undominated sequence at any time is consistent, then R and C are nondecreasing functions of P.

This property and its corollary yield a lower bound on the cost which can be used if a solution to the problem is known for a certain value of P smaller than the present one.

Consider variations in the duration of the mission T. Assume that $R(t)$ and C are the reliability function and cost of the optimal solution for a duration T, and $R'(t)$ and C' the reliability function and cost of the optimal solution for a duration T'. For simplicity, notational purpose take

$$Z(T)=\int_0^T R(t)dt \text{ and } Z(T')=\int_0^{T'} R'(t)dt.$$

THEOREM II: If $T'>T$, then $Z(T')-Z(T)\leq Z'(T')-Z'(T)$.

PROOF: By hypothesis, the following relationships hold:

$$PZ(T)-C\geq PZ'(T)-C'$$

$$PZ(T')-C\leq PZ'(T')-C'$$

These two inequalities imply that

$$Z(T)-Z'(T)\geq Z(T')-Z'(T') \text{ or } Z'(T')-Z'(T)\geq Z(T')-Z(T).$$

COROLLARY: If the undominated sequence (at a given time) is consistent, then $T'>T$ implies that $R'\geq R$ and $C'\geq C$.

This theorem and its corollary show that as T increases, the corresponding optimal solution is either a more costly and more reliable one in the case of consistency of the undominated sequence or a solution such that $\int Rdt$ becomes greater between T and the new horizon if the undominated sequence is not consistent.

THEOREM III: If the undominated sequence at a given time is consistent and if $P'\geq P$, then at the optimal solution, $m_i'^*\geq m_i^*$.

PROOF: Suppose that there are m_i^* components at stage i with P and $(m_i^*-v)=m_i'^*$ with P' (v integer >0). Let R^i and R'^i be the reliability on all stages, i excepted (i.e., R/R_i and R'/R_i') in the optimal solutions corresponding to P and P', respectively. Similarly let C^i and C'^i be the cost of these $(n-1)$ stages in the corresponding solutions.

PROOF: By the corollary of property I, $R'(t)\geq R(t)$ at the corresponding optimal solutions.

Thus, $R'^i>R^i$. Moreover, by hypothesis,

$$\text{(i)}\qquad P\int_0^T R_iR^idt-C^i-m_i^*c_i\geq P\int_0^T R_i'R^i-C^i-(m_i^*-V)c_i \text{ or } P\int_0^T (R_i-R_i')R^idt\geq vc_i.$$

Similarly,

(ii)

$$P' \int_0^T R_i' R'^i dt - C'^i - (m_i^* - v)c_i \geq P' \int_0^T R_i R'^i dt - C'^i - m_i^* c_i \text{ or } vc_i \geq P' \int_0^T (R_i - R_i') R'^i dt.$$

Grouping (i) and (ii), we get

$$P \int_0^T (R_i - R_i') R^i dt \geq P' \int_0^T (R_i - R_i') R'^i dt,$$

and, as $P < P'$ and $R_i(t) \geq R_i'(t)$, we get $R^i > R'^i$. Hence, a contradiction.

This theorem allows us to determine a minimum (maximum) number of components at each stage, if a solution is known for a smaller (larger) value of P and if the undominated sequences are consistent.

Generally, the reliabilities $R_i(m_i)$ are piecewise concave in m_i. To assume such a property is not a drastic restriction. In the case of loaded standbys, it is always satisfied. In the case of unloaded standbys, this is not always true. However, if we consider for example, components with exponential failure laws with mean $1/\lambda$, the gain in reliability by passing from m units to $(m+1)$ units is, for a one stage problem, $(\lambda t)^m e^{-\lambda t}/(m!)$. This gain is a decreasing function of m for $\lambda t < m$. Generally, problems will remain in such a situation because, if at the optimal solution m is less than λt, for a time t less than T, this implies that the reliability at the end is not very high. For $m = 1$, this implies a reliability smaller than e^{-1} at time T.*

If $R_i(m_i)$ is piecewise concave for all i's, that our assumption holds, we have the following properties.

Suppose that, for each stage, a maximum reliability attainable has been computed, i.e., at most $Z_i(t)$ the limit of $R_i(m_i, t)$ when m_i goes to infinity; this limit always exists because R_i is increasing in m_i and bounded by 1. Let R_M^i (M being used for representing the maximum) be the corresponding maximum reliability on all the stages, i excepted.

PROPERTY III: If the reliability at each stage is a concave function of the number of components at this stage, there exists an upper bound, M_i, on the number of components at each stage i such that

$$M_i = \min \left\{ k : P \int_0^T (R_i(k+1, t) - R_i(k, t)) R_M^i dt \leq C_i \right\}.$$

PROOF: The proof is analogous to that of Theorem V in [4].

Similarly, if we have a minimum number of components at each stage computed, and R_L^i is the minimum reliability on all the stages, i excepted, the following property holds:

PROPERTY IV: If the reliability at each stage is a concave function of the number of components at this stage, there exists a lower bound on the number of components at this stage, L_i, such that

$$L_i = \max \left\{ k : P \int_0^T (R_i(k, t) - R_i(k-1, t)) R_L^i dt \geq C_i \right\}.$$

These properties are also valid in the discounted case. A study of the asymptotic behavior of the function (1) yields the following propositions.

*We will show later that the gain on the integral of R_i; i.e., $(R_i(m_i, t) - R_i(m_i - 1, t))\, dt$, is a decreasing function of m_i and therefore, the following properties are also fully applicable to this situation.

THEOREM IV: If $P \to \infty$, then at the optimal solution, m_i^* tends to ∞ for all i.

PROOF: m_i^* tends to ∞, because L_i does. In order to have $L_i \geq N$ (N—arbitrary large), it is sufficient to take

$$P \geq C_i \Big/ \int_0^T (R_i(N, t) - R_i(N-1, t)) R_L^i dt,$$

where $R_L^i(t)$ can be taken as $R^i(1, \ldots, 1; t)$ to prove the theorem.

When T tends to infinity, results are less simple. However, it is possible to give some results depending upon the integrability of the reliability functions.

THEOREM V(a): If $T = \infty$, and if there exists an index j, such that $Z_j(t)$ be integrable on $[0, \infty]$, then at the optimal solution to (1), m_i^* remains finite for all i.

PROOF: We apply Lebesgue's theorem to the function

$$I(f(m_i)) = \int_0^\infty R_i(m_i, t) R_M^i dt;$$

$f(m_i)$ converges to $f = Z_i(t) R_M^i$ and is bounded by $Z_j(t)dt \in \mathscr{L}_1$. Then $I(f(m_i)) \in \mathscr{L}_1$ and converges to $I(f)$.

Therefore, $I(f(m_i)) - I(f(m_i - 1))$ converges to zero for m_i going to infinity. From Property III, there exists a maximum number of components M_i for stage i ($i = 1, \ldots, n$).

In the discounted case, it is easy to see that the above theorem is always applicable: it is sufficient to replace R_i by $R_i' = e^{-rt/n} R_i$ to see that $Z_i(t) \leq e^{-rt/n}$ is integrable on $[0, \infty]$. Therefore, in the discounted case, when the horizon is infinite, the optimal solution to the problem remains finite.

THEOREM V(b): If none of the Z_i belongs to $\mathscr{L}_1$, but if

$$\lim_{m_j \to \infty} \int_0^\infty (R_j(m_j, t) - R_j(m_j - 1, t)) L_M^j dt < C_j/P - \epsilon, \tag{3}$$

for ϵ positive number, then m_j^* remains finite when T goes to infinity.

PROOF: By contradiction, if we take $m_j^* \geq M$, with M large enough for

$$\int_0^\infty (R_j(M, t) - R_j(M-1, t)) L_M^j dt < C_j/P$$

we get $M \geq M_j$ and thus a contradiction.

For other situations, the asymptotic behavior of m_i^* is unknown. The function R may admit at least two maxima, one at finite range and the other at infinity. A priori, it is impossible to determine which one of these maxima would be the global one. However, if the left hand side of (2) is $> C_j/P + \epsilon$, making m_j going to infinity, increases the value of (1) to infinity.

4. AN EXAMPLE: UNLOADED STANDBYS

A situation which may frequently arise is the case of unloaded standbys. In this situation the standbys units have no probability of failure until they are put in the system. This may frequently

happen when only one stage is used; i.e., a system consists of one component and spare parts which are introduced in the system when the main component fails.

Let $q(t)$ be the unreliability of one of these components, $Q_n(t)$ the unreliability of the system made of one basic unit and $(n-1)$ standbys and let $A'_n(t)$ be the first derivative with respect to t of $Q_n(t)$. The unreliability of the system is given by the following expression [2];

$$Q_n(t) = \int_0^t q(t-\tau) Q'_{n-1}(\tau) d\tau \qquad \text{for } n > 1$$

$$Q_1(t) = q(t).$$

Such a law can be computed or approximated in most cases. For an exponential failure law, this formula becomes

$$Q_n(t) = 1 - \sum_{k=0}^{n-1} [(\lambda t)^k / k!]\, e^{-\lambda t}, \tag{4}$$

and (4) is easily integrable. The benefit of adding one unit to the system: i.e., the benefit of passing from n units to $(n+1)$ units is

$$P \int_0^T (\lambda t)^n / n!\, e^{-\lambda t} dt - c \qquad \text{(if } c \text{ is the cost of a component of this one stage system).}$$

By solving the integral, this function becomes

$$P\,[1 - e^{-\lambda T}(1 + \ldots + (\lambda T)^n / n!]/\lambda - c.$$

This function represents the profit of adding one unit to a one stage system with one basic unit and $(n-1)$ spare parts. It is decreasing in n and negative for n going to infinity. The solution n^* to our problem is therefore the smallest number n, such that the above expression is negative.

In the case of other failure laws, the integrals are not that easy to solve. If the unreliability of the unit $q(t)$ remains bounded between two linear functions of t: $\underline{\lambda} t \leq q(t) \leq \overline{\lambda} t$, the unreliability of the system is bounded between the unreliabilities of two components with exponential laws with parameters $\underline{\lambda}$ and $\overline{\lambda}$. Therefore, the above formulas applied to $\underline{\lambda}$ and $\overline{\lambda}$ give approximate solutions to the problem.

In the case of multi-stage problems, the above type of solution is not valid and the problem is more complex. However, for exponential failure laws, for example, all the terms in R, made of the products of expressions such as (4), can be integrated. If, for each stage, we compute a number n_i^*, as above, this number is not the optimal solution, but an upper bound on the number of components at each stage because the profit of adding a component is not

$$P \int_0^T (R_i(m_i, t) - R_i(m_i - 1, t))\, dt - c_i,$$

but

$$P \int_0^T (R_i(m_i, t) - R_i(m_i - 1, t))\, R^i dt - c_i. \tag{5}$$

If a lower bound is known on the number of components at each stage, (5) can be used, by replacing R^i by R^i to generate new lower bounds on the number of components at each stage.

5. COMPUTATIONAL PROCESSES

The former theorems, enable us to parametrize the problem as soon as an optimal solution is found for a value of P and T. However, the main difficulty remains the determination of such an optimal solution. We propose the following algorithm (in the case of R_i or $\int_0^T R_i dt$ concave in m_i):*

(a) Compute the limit for m going to infinity of $R_i\ (m, t)$ $(i=1, \ldots, n)$.

(b) Compute a maximum number of components at each stage (Property III). Stop the process when no further improvement is possible. Similarly, compute also the minimum number of components at each stage. If the minimum and maximum number of components at each stage coincide for each stage, stop. This is the solution. Otherwise, go to the next step.

(c) Select different times and compute the undominated sequence for each of these durations, the number of components at each stage remaining between the bounds determined at the former step. If these optimal sequences are identical, assume that the optimal sequence remains consistent (even if the system is not consistent, the solution will be very near the optimal one) and go to step *d*. If the optimal sequences remain identical, except for costs between a and b, assume that the optimal sequence is consistent for costs smaller than a and larger than b and go to step *f*. If the optimal sequences do not satisfy one of the two categories above, go to step *e*.**

(d) Compute the value of (1) for all the terms of the optimal sequence and terminate by taking the terms giving the largest value of (1). This is the optimal solution (or a solution very close to it).

(e) In this case, apply the branch and bound technique described in [3] and [4]. There are no fundamental, theoretical difficulties for its applications to the present problem; however, at each step, the integral of the reliability of the current solution considered in the algorithm must be computed, which lengthens the computations. The initial solution and lower bound on the function is either $(0, \ldots, 0)$ or $(M_1, \ldots, M_n)$, whichever one gives the largest value of (1).

(f) Compute the value of (1) for all the terms of the optimal sequence for costs smaller than a or larger than b. Take the term H among them which gives the largest value of (1). Then apply the branch and bound algorithm, as in step (3), with the following restrictions: all the complete solutions must have a cost between a and b and the initial solution and lower bound are H and the corresponding value of (1). The solution given by this algorithm is the optimal (or very close to the optimal) solution of the problem.

NOTE: It does not seem possible to use an approximate solution as we did in [4], because the cancellation of the first derivatives of the objective function does not give a solvable system of equations as in [1]. In some cases, however, (1) is also a concave function of the m_i. From [4], this is certainly true in the region of m_i's such that

$$\sum_{i=1}^{n} (1-R_i\ (m_i, T)) \leqslant 1,$$

*As mentioned before, in the discounted case, R_i is replaced by $e^{-rt/n}R_i$.

** In the case where the undominated sequence would remain consistent except on some intervals $[a_j, \ldots, b_j]$, it would be possible to extend the method described in *f*.

or everywhere if we have loaded standbys. If $(M_1, \ldots, M_n)$ satisfy such a condition, it seems possible to find a local maximum (which would probably be a global one too) by applying the same routine described in [4]. Now, if this solution did not satisfy (4), we would have to show that the function (1) is still concave at that point or forget it and apply the algorithm described above.

REFERENCES

[1] Fan, L. T., C. S. Wang, F. A. Tillman, and Huang, "Optimization of System Reliability by the Discrete Maximum Principle," IEEE Transactions on Reliability, *16* (Sept. 1967).

[2] Gnedenko, B. V., Y. K. Belyayev, and A. D. Solovyev, *Mathematical Methods of Reliability Theory* (Academic Press, 1969).

[3] Henin, C. G., "An Algorithm for Maximizing Reliability Through System Redundancy," Carnegie-Mellon University, Management Sciences Report #216 (Aug. 1970).

[4] Henin, C. G., "Computational Techniques for Optimizing Systems with Standby Redundancy," Nav. Res. Log. Quart. *19*, 293–308 (June 1972).

[5] Ketelle, J. D., "Least Cost Allocation of Reliability Investment," Operations Research, *10*, 249–265 (1962).

A CONTINUOUS SUBMARINE VERSUS SUBMARINE GAME

Eric Langford*

University of Maine
Orono, Maine

ABSTRACT

This paper analyzes, from a game-theoretic standpoint, the simultaneous choice of speeds by a transitor and by an SSK which patrols back and forth perpendicular to the transitor's course. Using idealized acoustic assumptions and a cookie-cutter detection model which ignores counterdetection, we are able to present the problem as a continuous game, and to determine an analytic solution. The results indicate that with these assumptions, there are conditions under which neither a "go fast" nor a "go slow" strategy is optimal. The game provides a good example of a continuous game with a nontrivial solution which can be solved effectively.

INTRODUCTION

This paper analyzes from a game-theoretic standpoint the simultaneous choice of speeds by a transitor and by an *SSK* which patrols back and forth perpendicular to the transitor's course. (An *SSK* is a submarine whose mission is directed against enemy submarines.) The payoff, in effect, is taken to be the *SSK*'s detection sweep width. This game was originally investigated by D. H. Wagner and E. P. Loane in classified reports during 1963–64. Their treatment was confined to choices of speeds from a discrete set, but applied to rather general acoustic conditions. The present analysis assumes an idealized form of propagation loss versus range and of noise versus speed. These idealizations permit the sweep width to be expressed as a convenient continuous function of the two speeds; accordingly, they allow each player to make choices from a continuum of speeds, rather than from a discrete set. They also permit a comprehensive analysis of the variety of cases which can arise within the idealizations.

Subsequent to Wagner and Loane's work, an approach to this game was undertaken by Mathematica [6], treating a continuous analytic payoff function based on idealized acoustics. Unfortunately, inconsistent acoustic assumptions were used in [6]; these were corrected by Mathematica in a subsequent report [1]. Motivated by [6] (and prior to [1]), the present analysis was undertaken, again using a continuous payoff function, but using acoustic assumptions generally felt to be consistent.

In this paper, we assume that the graph of noise versus speed is linear above a breakpoint speed below which noise is independent of speed; this is a common assumption and was made in [6]. We also make the usual "spreading law" assumption, i.e., that propagation loss is proportional to k'th power of range when loss is expressed in power units. This of course is equivalent to the assumption that propagation loss is proportional to the logarithm of the (k'th power of) range when loss is expressed in decibels. (The inconsistency in [6] appears to be on this point.) We take for a payoff function what is essentially the *SSK*'s kinematically enhanced sweep width; the *SSK*, of course, attempts to

---

*Research on this paper was performed when the author was with the Naval Postgraduate School and Daniel H. Wagner, Associates.

maximize this quantity and the transitor attempts to minimize it. The *SSK* will thus be maximizing his detection probability if we assume a cookie-cutter detection model, i.e., that detection occurs when and only when the signal excess (assumed deterministic) reaches a threshold value. We ignore counter-detection by the transitor. This will be a realistic assumption if the *SSK*'s acoustic capability is far superior to that of the transitor. The more general problem, which takes into account counterdetection and which allows for a stochastic detection model, appears to be quite difficult to solve within this framework. However, in a discretized form, it was handled satisfactorily by Wagner and Loane in the above-mentioned reports.

The game is described in abstract terms in the first section, i.e., the payoff function is given in a form which is equivalent (for purposes of the game's analysis) to the formula for the *SSK*'s kinematically enhanced sweep width. (The *SSK*'s effective sweep width is increased by his back and forth patrol.) The properties of the payoff function are developed into geometric criteria for solution. In the second section, graphical methods of finding the solution are given and illustrated by examples. A numerical solution is described in the third section. The fourth section enumerates the possible outcomes as combinations of the pure and mixed strategies. Identification of the abstract game with the real *SSK* versus transitor game is given in the final section.

In an earlier version, this paper was submitted as a Memorandum to Commander, Submarine Development Group Two, in New London [4]. In this earlier version are included graphs of all possible cases and a Fortran computer program to solve the game. A subsequent classified memorandum applied the analysis to some "real life" numbers.

DESCRIPTION OF THE GAME

We consider the following game. The maximizing player (*SSK*) chooses a speed u in the range $0 \leq u_{\min} \leq u \leq u_{\max}$, and the minimizing player (transitor) chooses a speed v in the range $0 < v_{\min} \leq v \leq v_{\max}$. The payoff function is

$$F(u, v) = e^{cv - u} \sqrt{1 + u^2 / v^2},$$

where $c > 0$ is a constant.

The explicit identification of this payoff function with the *SSK*'s detection sweep width will be clarified later. For the time being, we treat the game abstractly.

Computing the partial derivative with respect to v, we see that

$$F_2(u, v) = F(u, v)\left[c - \frac{u^2}{v(u^2+v^2)}\right].$$

The second partial derivative with respect to v is

$$F_{22}(u, v) = F(u, v)\left\{\left[c - \frac{u^2}{v(u^2+v^2)}\right]^2 + \frac{u^2(u^2+3v^2)}{v^2(u^2+v^2)}\right\},$$

so that $F_{22} > 0$. That is, F is convex in its second argument; it is well-known that this implies the following facts (see [2, p. 80] or [5, p. 259]):

1. The minimizing player (transistor) always has an optimal pure strategy.

2. The maximizing player (*SSK*) always has an optimal mixed strategy involving at most two speed choices; i.e., he either has an optimal pure strategy, or an optimal mix of two speeds.

Let us fix v and consider F as a function of the one variable u. Since F is continuous and restricted to a closed bounded interval, it assumes a maximum value; this maximum must be taken on at an endpoint (i.e., either $u_{\min}$ or $u_{\max}$) or at an interior relative maximum, which can, in this case, be located by differentiation. To this end, we form the partial derivative of F with respect to u, and equate it to zero:

$$F_1(u, v) = F(u, v)\left[\frac{u}{u^2 + v^2} - 1\right] = 0.$$

Now F never vanishes, so that $F_1(u, v) = 0$ iff $u = u^2 + v^2$. The solution set to this equation in the $u-v$ plane is the semicircle

$$\{(u, v): (u - 1/2)^2 + v^2 = 1/4, \text{ and } v > 0\}.$$

From the form of F_1, it is evident that $F_1(u, v) > 0$ if (u, v) is inside the semicircle, and that $F_1(u, v) < 0$ if (u, v) is outside the semicircle. Thus the graph of $F(u, v)$ versus u for fixed v can have three qualitatively different forms as shown in Figure 1. The case $v = 0.6$ is typical of $v > 1/2$ and the case $v = 0.4$ is typical of $v < 1/2$.

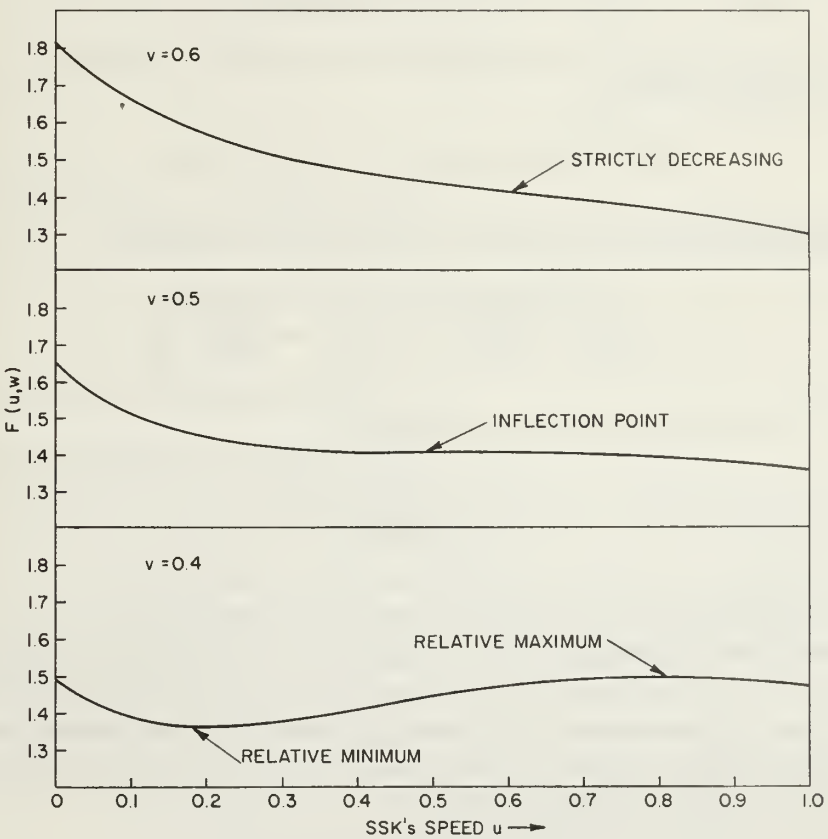

FIGURE 1. Qualitatively different forms of F.

Let us define the function φ as follows:

$$\varphi(v) = \max \{F(u, v): u_{\min} \leqslant u \leqslant u_{\max}\}.$$

From the above observations, it is not difficult to infer the following:

(1) If $v \geqslant 1/2$, then $\varphi(v) = F(u_{\min}, v)$.

(2) If $v < 1/2$, then

$$\varphi(v) = \max \{F(u_{\min}, v), F(u_{\max}, v), F(u_0, v)\}, \tag{1}$$

where $u_0 = \frac{1}{2} + \sqrt{\frac{1}{4} - v^2}$, if this falls within the interval $u_{\min} \leqslant u_0 \leqslant u_{\max}$; otherwise we ignore u_0.

Since F is convex in its second argument, v, it follows that φ is continuous. Its minimum value will be the value of the game; moreover, if φ assumes its minimum at $v = v_0$, then v_0 is an optimal pure strategy for the minimizing player. By the convexity of F, φ is unimodal (hence the above v_0 is defined uniquely). Since φ is unimodal, we can locate its minimum numerically with great efficiency using a binary search; graphical methods are also possible.

Similarly, we can fix u and solve the equation $F_2(u, v) = 0$:

$$F_2(u,v) = F(u,v)\left[c - \frac{u^2}{v(u^2+v^2)}\right] = 0.$$

The solution set to this equation is

$$\{(u, v): cv^3 + cu^2v = u^2, \qquad (u, v) \neq (0, 0)\}.$$

By the Cardan-Tartaglia formula, this defines v as a function of u as follows:

$$v = \sqrt[3]{\frac{u^2}{2c} + u^2\sqrt{\frac{1}{4c^2} + \frac{u^2}{27}}} + \sqrt[3]{\frac{u^2}{2c} - u^2\sqrt{\frac{1}{4c^2} + \frac{u^2}{27}}}.$$

Conversely, if $v < 1/c$, we can solve for u as a function of v:

$$u = \sqrt{\frac{cv^3}{1 - cv}}.$$

Figure 2 shows the graphs of $F_1(u, v) = 0$ and $F_2(u, v) = 0$ for several values of c.

The intersection of the graphs of $F_1(u, v) = 0$ and $F_2(u, v) = 0$ will give a saddle point if $u > 1/2$, i.e., if $c > 1$. If there is a pure strategy minimax solution to the game which is interior for both players (i.e., neither $u_{\min}$ nor $u_{\max}$ for the *SSK* and neither $v_{\min}$ nor $v_{\max}$ for the transitor), then the solution must occur at this saddle point. (Edge minimaxes need not be of this form.) To find this saddle point, we solve the following equations simultaneously:

$$F_1(u, v) = F_2(u, v) = 0.$$

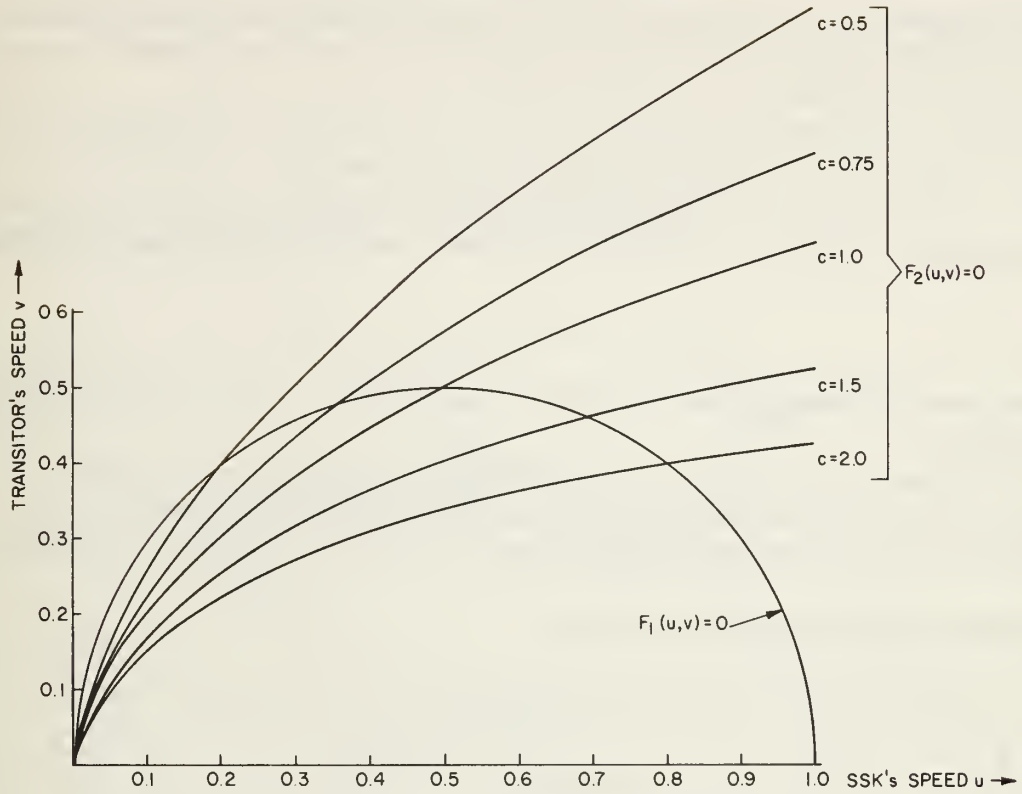

FIGURE 2. Graphs of $F_1(u, v)=0$ and of $F_2(u, v)=0$ for several values of c.

Strangely enough, the answer is exceedingly simple:

$$u=\frac{c^2}{c^2+1} \qquad v=\frac{c}{c^2+1}.$$

Note that there can be no interior minimax if $c \leq 1$.

GRAPHICAL SOLUTION OF THE GAME

Define the function f as follows: For any u_0, $u_{\min} \leq u_0 \leq u_{\max}$, we define $v_0=f(u_0)$ iff $F(u_0, v_0)=\min\{F(u_0, v): v_{\min} \leq v \leq v_{\max}\}$. That is, for any admissible u, $f(u)$ is the v which minimizes F. By the v-convexity of F, f is a well-defined, single-valued function. Moreover, f is continuous and monotone increasing. In fact, if we stay away from $v_{\min}$ and $v_{\max}$, f is strictly increasing; more precisely, if $u_1 < u_2$ are such that $v_{\min} < f(u_1) \leq f(u_2) < v_{\max}$, then f is strictly increasing on the interval $[u_1, u_2]$. See Figures 5, 6, and 7 for examples.

Actually, we can give a simple formula for f. Let $h(u)$ be the solution to $F_2(u, h(u))=0$, i.e.

$$h(u)=\sqrt[3]{\frac{u^2}{2c}+u^2\sqrt{\frac{1}{4c^2}+\frac{u^2}{27}}}+\sqrt[3]{\frac{u^2}{2c}-u^2\sqrt{\frac{1}{4c^2}+\frac{u^2}{27}}}.$$

For unrestricted v, $F(u_0, v)$ is convex and has an absolute minimum at $v = h(u_0)$. Therefore if $v_{\min} \leq h(u_0) \leq v_{\max}$, then $f(u_0) = h(u_0)$. If $h(u_0) \geq v_{\max}$, then $f(u_0) = v_{\max}$, and if $h(u_0) \leq v_{\min}$, then $f(u_0) = v_{\min}$. We can summarize these cases as follows:

$$f(u_0) = \min \{v_{\max}, \max [v_{\min}, h(u_0)]\}.$$

It is straightforward to determine $f(u_0)$ graphically. Let u_0 be fixed and consider the line segment

$$S = \{(u, v): u = u_0 \text{ and } v_{\min} \leq v \leq v_{\max}\}.$$

The three possibilities, namely $f(u_0) = v_{\min}$, $f(u_0) = v_{\max}$, and $f(u_0) = h(u_0)$ can be exhibited graphically as follows:

(1) If the line S lies completely above the graph of $h(u)$ versus u, then $f(u_0) = v_{\min}$.

(2) If the line S lies completely below the graph of $h(u)$ versus u, then $f(u_0) = v_{\max}$.

(3) If the line S intersects the graph of $h(u)$ versus u at (u_0, v_0), then $f(u_0) = v_0 = h(u_0)$.

These three possibilities are graphed in Figure 3.

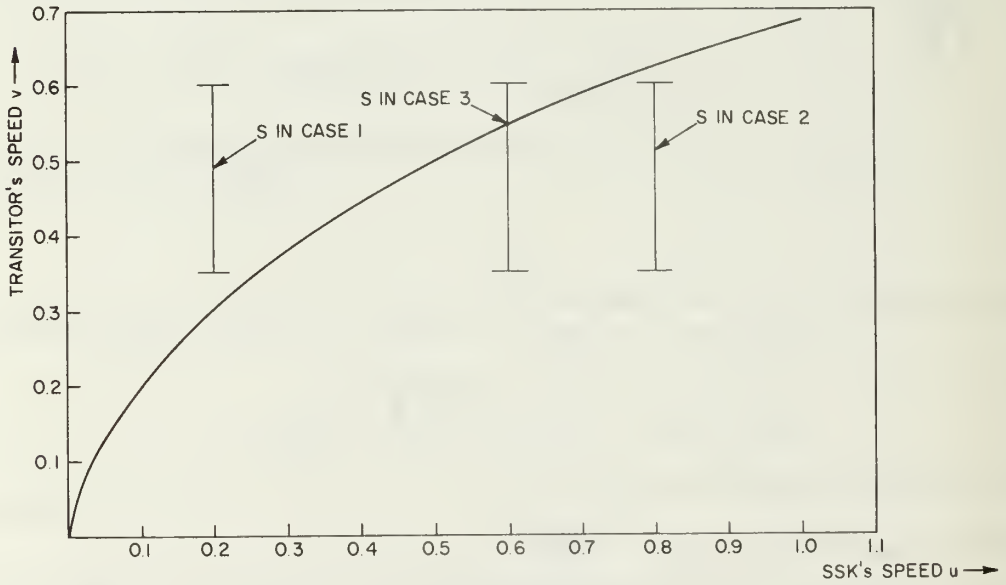

FIGURE 3. Three possibilities in the determination of $f(u)$; ($c = 1$ in this example).

In a similar fashion we define the "function" g as follows: For any v_0, $v_{\min} \leq v_0 \leq v_{\max}$, we define $u_0 = g(v_0)$ iff $F(u_0, v_0) = \max \{F(u, v_0): u_{\min} \leq u \leq u_{\max}\} = \varphi(v_0)$. From Figure 1, it is clear that this need not, in general, define a single-valued function. We shall subsequently show that either g is a continuous single-valued function or that the graph of g consists of two continuous pieces which overlap only at an endpoint. More precisely, in this second circumstance, there exists a point v_0 such that $v_{\min} \leq v_0 \leq v_{\max}$, and such that g restricted to either of the sub-intervals $[v_{\min}, v_0)$ or $(v_0, v_{\max}]$ is a continuous single-valued function; however at $v = v_0$, $g(v)$ is bivalent; it has the two values $g(v_0 - 0)$ and $g(v_0 + 0)$. (It will also be shown later that this right-hand limit can only be $u_{\min}$ and that the left-

hand limit can be either $u_{\max}$ or $u_0 = 1/2 + \sqrt{1/4 - v_0^2}$.) In either case, g is monotone decreasing: $g(v_1) \geq g(v_2)$ whenever $v_1 \leq v_2$. Figure 5 illustrates the first possibility, while Figures 6 and 7 illustrate the second.

Note that when we fix v and regard $F(u, v)$ as a function only of u, the constant c enters the expression for $F(u, v)$ only as a constant multiplier. Since we are interested in locating the maximizing u for fixed v, it follows that the value of c is unimportant in the discussion which follows. That is, the maximizing u is located in the same place no matter what value c has.

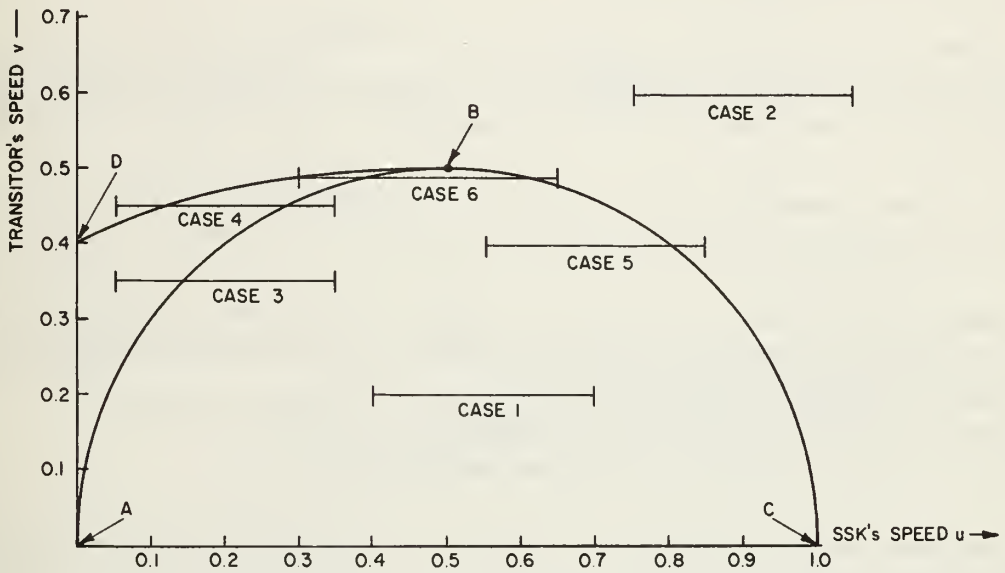

FIGURE 4. Six possibilities in the determination of $g(v)$.

We refer to Figure 4: the semicircle ABC is the locus of $F_1(u, v) = 0$. The points on the quarter-circle AB are relative minima for $F(u, v)$ considered as a function of u for fixed v, and the points on the quarter-circle BC are relative maxima. The point B itself is an inflection point (cf. Figure 1). The curve DB is defined as follows: if (u_1, v_0) lies on DB, then $F(u_1, v_0) = F(u_2, v_0)$, where (u_2, v_0) lies on the quarter-circle BC. The curve DB is thus obtained as the solution set of the following transcendental equation

$$F(u, v) = F(1/2 + \sqrt{1/4 - v^2}, v),$$

where $0 \leq u \leq 1/2$.

Let v_0 be fixed and let S' denote the line segment

$$\{(u, v): u_{\min} \leq u \leq u_{\max} \text{ and } v = v_0\}.$$

There are six possibilities:

(1) The line S' lies completely within ABC; then $g(v_0) = u_{\max}$. The line S' may meet, but not cross ABC.

(2) The line S' lies completely outside of ABC; then $g(v_0) = u_{\min}$. The line S' may meet, but not cross ABC.

(3) The line S' crosses AB, but does not meet or cross DB. In this case, $g(v_0)$ may be $u_{\min}$, $u_{\max}$, or both, depending on the relative sizes of $F(u_{\min}, v_0)$ and $F(u_{\max}, v_0)$.

(4) The line S' crosses DB; then $g(v_0) = u_{\min}$.

(5) The line S' meets or crosses BC, but does not meet DB. In this case, the maximizing u_0 occurs at the intersection of S' and BC. Evidently $g(v_0) = u_0 = \frac{1}{2} + \sqrt{\frac{1}{4} - v_0^2}$.

(6) The line S' meets, but does not cross DB. In this case, $g(v_0) = u_{\min}$, *unless* S' also meets or crosses BC. In this latter circumstance, $g(v_0)$ has the two values $u_{\min}$ and u_0, where u_0 is at the intersection of S' and BC as in the case above.

These possible cases are all graphed in Figure 4.

If we graph $g(v)$ versus v, the graph will be a nice continuous curve as long as $g(v)$ "stays put", i.e. as long as $g(v)$ is one or the other of the endpoints or is u_0. Problems arise at the two "transition stages":

A. As in Case 3, when $F(u_{\min}, v_0) = F(u_{\max}, v_0) = \varphi(v_0)$. This will be called a type A transition.

B. As in Case 6, when $F(u_{\min}, v_0) = F(u_0, v_0) = \varphi(v_0)$. This will be called a type B transition.

As v increases through a transition stage v_0, $g(v)$ will make an abrupt jump. Precisely *at* the transition stage $v = v_0$, $g(v)$ will be two-valued. It will now be shown that as v increases through the transition stage v_0, one of the following two things will happen:

A. If v_0 is a type A transition, then $g(v_0)$ must jump from $u_{\max}$ to $u_{\min}$.

B. If v_0 is a type B transition, then $g(v_0)$ must jump from u_0 to $u_{\min}$.

It is a consequence of the above that g can have at most one transition stage, since any transition puts $g(v)$ in the "absorbing state" $u_{\min}$. An example of a type A transition is found in Figure 7, and an example of a type B transition is found in Figure 6.

To prove the above assertion, let us suppose that $u_1 < u_2$ and that for some v_0, $F(u_1, v_0) = F(u_2, v_0)$. If we define the function $G(v) = F(u_1, v) - F(u_2, v)$, then by assumption, $G(v_0) = 0$. By computing the (continuous) derivative $G'(v)$, it is easy to show that $G'(v_0) > 0$, so that G is increasing in a neighborhood of v_0; in other words, as v increases through v_0, $F(u_1, v)$ is first less than $F(u_2, v)$ and then greater, so that a transition can occur from a larger value of u to a smaller value, never the other way.

The "function" g is monotone decreasing since it decreases at its transition point (if there is one) and since it obviously must decrease at points other than transition points. As is the case with f, if $v_1 < v_2$ are such that $u_{\min} < g(v_2) \leq g(v_1) < u_{\max}$, then g is strictly decreasing on the interval $[v_1, v_2]$.

If we plot both f and g within the rectangle of admissible speeds defined by

$$\{(u, v) : u_{\min} \leq u \leq u_{\max} \text{ and } v_{\min} \leq v \leq v_{\max}\},$$

then one of two things will occur:

(1) The two graphs will intersect at a single point (u_0, v_0). In this case, u_0 is an optimal pure strategy for the maximizing player (*SSK*) and v_0 is an optimal pure strategy for the minimizing player (transitor). See Figure 5 for an example of this case.

(2) The two graphs will not intersect. More precisely, there will exist a transition point v_0 such that $g(v_0)$ has the two values u_1 and u_2, where $u_1 < u_2$ and where $u_1 < f^{-1}(v_0) < u_2$. See Figures 6 and 7 for examples of this. Note that f^{-1} is defined at v_0 since f is strictly increasing as it passes through

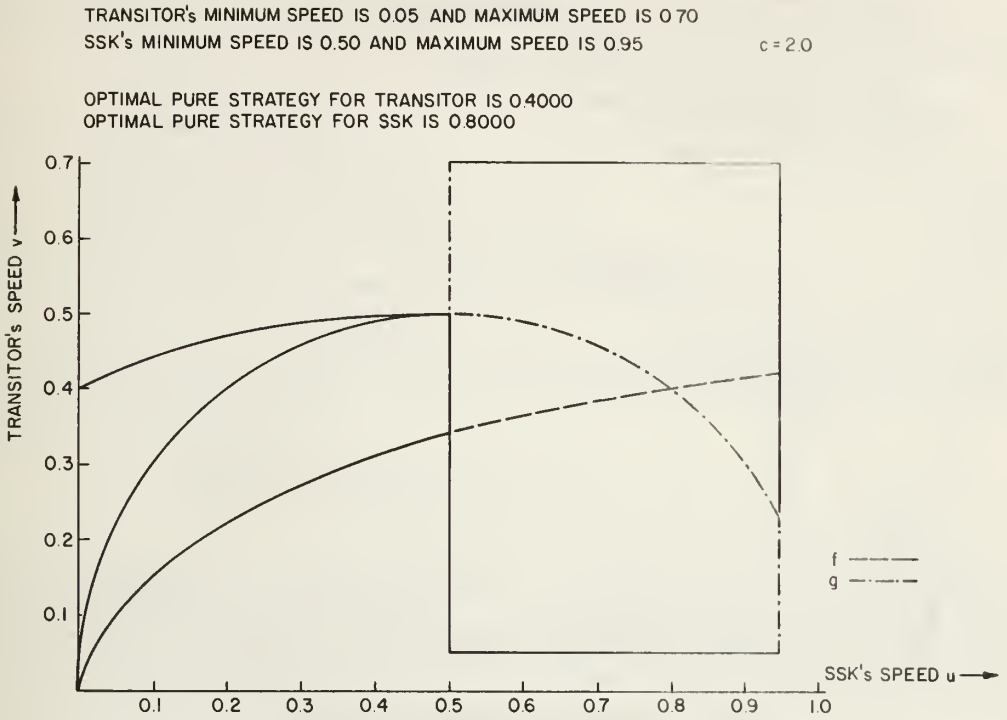

FIGURE 5. Graphical solution of the game—Example 1.

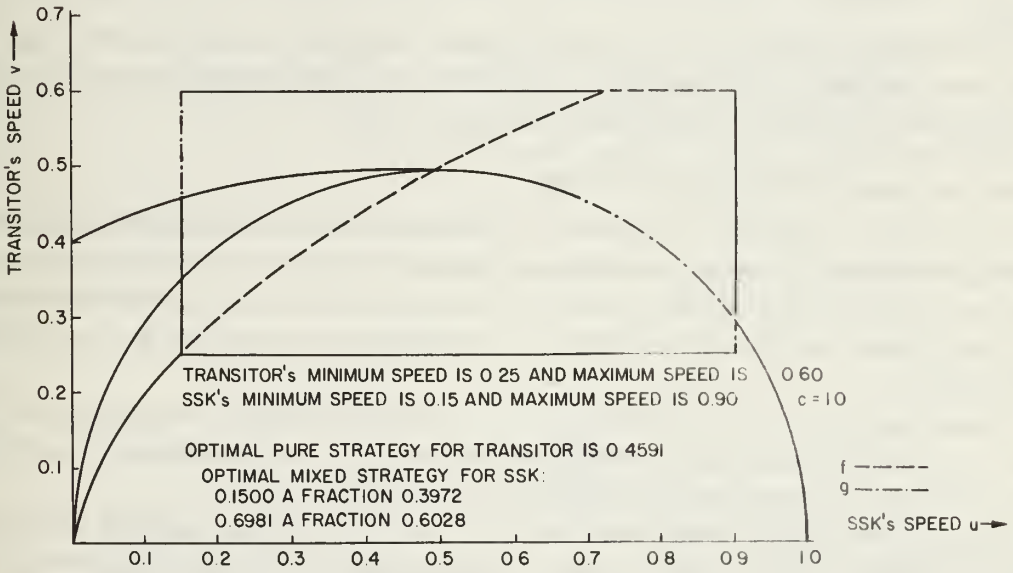

FIGURE 6. Graphical solution of the game—Example 2.

the hole in the graph of g. In this case, the minimizing player again has the optimal pure strategy v_0, but the maximizing player now has an optimal mixed strategy given by u_1 with probability p and u_2 with probability $1-p$. The constant p is found by solving the following equation:

$$p \cdot F_2(u_1, v_0) + (1-p) \cdot F_2(u_2, v_0) = 0.$$

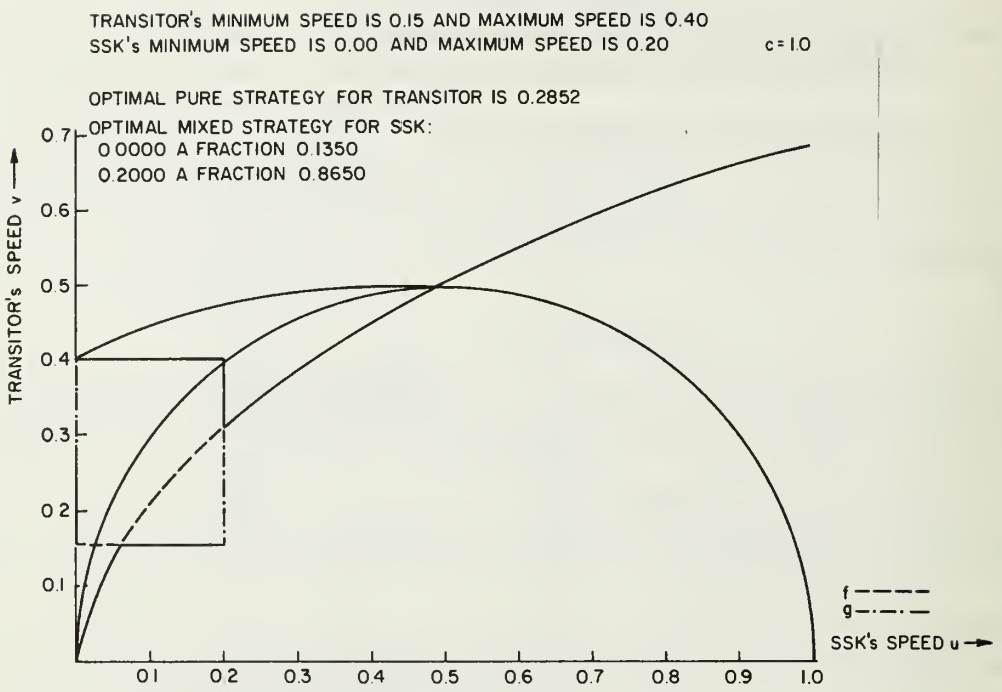

FIGURE 7. Graphical solution of the game—Example 3.

No other possibilities (e.g., two intersections) can arise by the aforementioned monotonicity properties of f and g.

Three numerical examples of this graphical solution are given to make this more clear. In each of these figures, the graph of f is indicated by dash line and the graph of g is indicated by a dot-dash line. The box is the rectangle of admissible speeds defined above.

NUMERICAL SOLUTION OF THE GAME

The following is a step-by-step procedure for solving the game. It essentially repeats the graphical procedure, but from a point of view which emphasizes suitability for computation. This procedure has been programmed for the GE–235 computer: a copy of the program and sample output are included in [4]. The following step-by-step procedure can be considered a macroscopic flow chart of the computer program.

(1) Locate the minimum of $\varphi(v) = \max_u F(u, v)$. Since φ is unimodal, this can be done easily by iterative computation using a binary search. The function $\varphi(v)$ itself is evaluated by using formula (1) derived earlier. The minimum value of $\varphi(v)$ is the value of the game.

(2) Let v_0 be that unique number such that $\varphi(v_0) = \min_v \varphi(v)$. This is obtained automatically as a by-product of step 1. Then v_0 is the optimal pure strategy for the minimizing player (transitor).

(3) Solve the equation $F(u, v_0) = \varphi(v_0)$ for u by checking the three possible places where the maximum could occur (viz. $u_{\min}$, $u_{\max}$, and $u_0 = \frac{1}{2} + \sqrt{\frac{1}{4} - v_0^2}$). If there is a unique solution u^*, then u^* is an optimal pure strategy for the maximizing player (*SSK*). If there are two distinct solutions $u_1 < u_2$, then a mix of u_1 with probability p and u_2 with probability $1-p$ will be optimal. The number p is the solution to the equation

$$p \cdot F_2(u_1, v_0) + (1-p) \cdot F_2(u_2, v_0) = 0.$$

ENUMERATION OF POSSIBLE OUTCOMES

From an examination of Figure 1 or otherwise, we see that there are five qualitatively distinct possibilities for the maximizing player:

$U(1)$. A pure strategy of $u_{\min}$.

$U(2)$. A pure strategy of $u_{\max}$.

$U(3)$. A pure strategy of u_0, where $u_{\min} < u_0 < u_{\max}$.

$U(4)$. A mixed strategy of $u_{\min}$ and $u_{\max}$.

$U(5)$. A mixed strategy of $u_{\min}$ and u_0, where $u_{\min} < u_0 < u_{\max}$.

Since the minimizing player always has a pure strategy, there are only three possibilities for him:

$V(1)$. A pure strategy of $v_{\min}$.

$V(2)$. A pure strategy of $v_{\max}$.

$V(3)$. A pure strategy of v_0, where $v_{\min} < v_0 < v_{\max}$.

Apparently there are 15 cases to consider. However it is known [5, p. 267] that in cases $V(1)$ and $V(2)$, the maximizing player must also have a pure strategy. (This can be inferred also from the geometric reasoning previously used.) Thus there are only 11 possible cases. In [4], examples are given of each of these 11 cases. We remark that for certain values of c, some cases are forbidden. For example, if $c \leq 1$, then the case $U(3) - V(3)$ cannot occur.

IDENTIFICATION OF ABSTRACT GAME WITH REAL GAME

We shall now identify the foregoing abstract game with a real *SSK* versus transitor game.

The *SSK* moves back and forth at constant speed u' across a rectangular barrier zone. The transitor enters the zone on a course perpendicular to the *SSK*'s course at a constant speed v'.

The payoff for the *SSK* is detection sweep width. He attempts to maximize this quantity and the transistor attempts to minimize it. Let $L_S(v')$ denote the radiated noise (in decibels measured at 1 yard from the source) of the transitor as a function of its speed v', and let $L_N(u')$ denote the self noise (in decibels) of the *SSK* as a function of its speed u'.

We assume that the noise curves of both *SSK* and transitor are of the form shown in Figures 8 and 9, respectively.

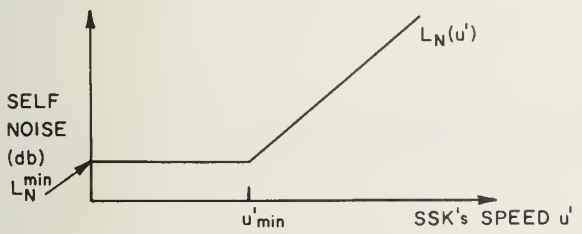

FIGURE 8. Noise curve for *SSK*.

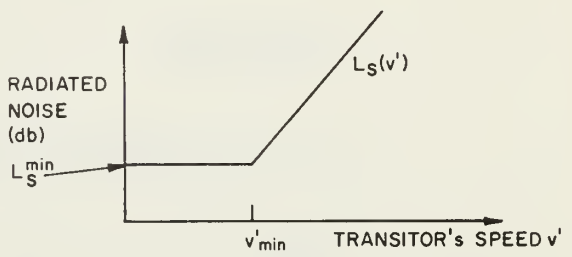

FIGURE 9. Noise curve for transitor.

It is obvious that the *SSK* will never travel more slowly than $u'_{\min}$ and that the transitor will never travel more slowly than $v'_{\min}$. We therefore assume that $L_N(u')$ and $L_S(v')$ are both linear for *all* u' and v', but in the analysis we will not consider any speeds less than these minimum speeds. The

maximum speeds, $u'_{\max}$ and $v'_{\max}$, are of course determined by the physical characteristics of the respective submarines. We have then the following formulas for $L_N(u')$ and $L_S(v')$:

$$L_N(u') = L_N^{\min} + b(u' - u'_{\min})$$

$$L_S(v') = L_S^{\min} + a(v' - v'_{\min}),$$

where a and b are the slopes, measured in decibels/knot.

Assume that propagation loss obeys a spreading law, so that the decibel loss in propagating from 1 to R yd is $k \log_{10} R$, for some fixed $k > 0$.

The unenhanced sweep width W for the *SSK* is given by

$$k \log_{10}(W/2) = L_S(v') - L_N(u') + N_{DI} - N_{RD},$$

where N_{DI} and N_{RD} are the *SSK*'s directivity index and recognition differential. (Of course all term in this equation are taken at the frequency and bandwidth of interest.)

If we solve this equation for W, we obtain

$$W = 2 \exp\left\{\frac{1}{k}(\log 10)[L_S(v') - L_N(u') + N_{DI} - N_{RD}]\right\}$$

$$= 2 \exp\left\{\frac{1}{k}(\log 10)[L_S^{\min} + av' - av'_{\min} - L_N^{\min} - bu' + bu'_{\min} + N_{DI} - N_{RD}]\right\}$$

$$= 2 \exp\left\{\frac{1}{k}(\log 10)[L_S^{\min} - av'_{\min} - L_N^{\min} + bu'_{\min} + N_{DI} - N_{RD}]\right\} \cdot \exp\left\{\frac{b}{k}(\log 10)\left[\left(\frac{a}{b}\right)v' - u'\right]\right\}.$$

If we make the following substitutions:

$$u = \left(\frac{b}{k} \log 10\right) u'$$

$$v = \left(\frac{b}{k} \log 10\right) v'$$

$$K = 2 \exp\left\{\frac{1}{k}(\log 10)[L_S^{\min} - L_N^{\min} - av'_{\min} + bu'_{\min} + N_{DI} - N_{RD}]\right\}$$

$$c = \frac{a}{b},$$

then W is of the form

$$W = Ke^{cv-u},$$

where K and c are positive constants.

As noted by Wagner and Loane and by Mathematica, the kinematic enhancement of sweep width in the back and forth patrol is a multiplicative increase in the approximate amount $\sqrt{1+(u'/v')^2}$, providing the *SSK*'s patrol legs are substantially longer than the acoustic sweep width W. (See [3, Equation (7.1.4.)].) Since the kinematic enhancement factor depends only on the ratio $u'/v' = u/v$, the results of the previous analysis apply. Here the primed variables, namely u' and v', refer to true ship speeds (in knots). The unprimed variables, namely u and v, refer to the "normalized speeds" as considered in the solution to the game. All graphs, etc. are referred to these normalized speeds.

ACKNOWLEDGMENTS

I would like to thank Dr. Daniel H. Wagner for his help in the preparation of this paper. The work was originally supported by ONR Contract Nonr–4784(00). The writing of this paper was supported by an ONR Foundation Grant (FY1968).

REFERENCES

[1] Agin, Norman I., et al., "The Application of Game Theory to ASW Detection Problems," Mathematica Report, Princeton, New Jersey (Sept. 30, 1967).

[2] Karlin, S., *Mathematical Methods and Theory in Games, Programming, and Economics* (Addison-Wesley, Reading, Mass., 1959).

[3] Koopman, B. O., "Search and Screening," OEG Report No. 56, Operations Evaluation Group, Office of the Chief of Naval Operations, Navy Department (1946).

[4] Langford, E. S., "Game–Theoretic Analysis of Choice of Speeds by SSK and Transitor," Daniel H. Wagner Associates Memorandum to CSDG–2 (Nov. 17, 1966).

[5] McKinsey, J., *Introduction to the Theory of Games* (McGraw-Hill, New York, 1952).

[6] "A Study of Optimal Patrol and Transit Strategies in a Rectangular Barrier Zone Using Mathematical Games," Mathematica Report, Princeton, New Jersey.

TOTAL OPTIMALITY OF INCREMENTALLY OPTIMAL ALLOCATIONS*

Lawrence D. Stone

Daniel H. Wagner, Associates

Paoli, Pennsylvania

ABSTRACT

This paper considers the problem of finding optimal solutions to a class of separable constrained extremal problems involving nonlinear functionals. The results are proved for rather general situations, but they may be easily stated for the case of search for a stationary object whose a priori location distribution is given by a density function on R, a subset of Euclidean n-space. The functional to be optimized in this case is the probability of detection and the constraint is on the amount of effort to be used.

Suppose that a search of the above type is conducted in such a manner as to produce the maximum increase in probability of detection for each increment of effort added to the search. Then under very weak assumptions, it is proven that this search will produce an optimal allocation of the total effort involved. Under some additional assumptions, it is shown that any amount of search effort may be allocated in an optimal fashion.

1. INTRODUCTION

In this paper we consider the relationship between incrementally optimal allocations and totally optimal allocations. Motivation for studying this relationship arises naturally in planning a search for a lost object. Suppose that the search planner is given authorization to search for a fixed time interval, and he conducts the search to produce the maximum probability of detection at the end of the interval. If the search fails to detect the lost object within the allotted time, the planner may be given authorization to continue searching for an additional time increment. In this case the planner may allocate the additional search effort to maximize the probability of detection in the given increment. Having done this, one may ask whether the search could have produced a higher detection probability if it were known in advance that both the initial time interval and the added increment were available.

In mathematical terms the search problem is to allocate optimally a given amount of effort in order to detect a stationary object, the target, located in Euclidean n-space, R. There is a function f which gives the probability density of the target's location. Suppose T is the amount of effort available for the search. Then the search planner seeks a function $q^*: R \to [0, \infty)$ such that $\int_R q^*(x)dx \leq T$ and

$$\int_R b(x, q^*(x)) f(x)dx = \max\left\{\int_R b(x, q(x)) f(x)dx : q \geq 0, \int_R q(x)dx \leq T\right\}. \tag{1.1}$$

The function $b(x, \cdot)$ is the *local effectiveness function* at x. That is, $b(x, y)$ gives the conditional probability of detecting the target given it is located at x and the effort density is y at x. The integral on the left of (1.1) gives the probability of detecting the target when using allocation q^*. The function q^* is called an optimal allocation. This problem has an obvious analog when R is replaced by a countable set of locations or boxes.

*This research was supported by the Naval Analysis Programs, Office of Naval Research, under Contract No. N00014-69-C-0435.

For the case where $b(x, y) = 1 - e^{-y}$ for $x \in R$ and $y \geq 0$, Koopman [4, p. 617] made the following observation. Suppose one allocates T_1 amount of effort in an optimal fashion, but fails to detect the target. An increment T_2 of effort then becomes available. If one allocates this additional effort in an incrementally optimal manner (i.e., optimal considering the previous allocation of T_1 amount of effort), then one obtains an optimal allocation of $T_1 + T_2$ effort. That is, two incrementally optimal allocations produce a totally optimal allocation.

In [2] an incomplete attempt was made to show that incrementally optimal allocations produce totally optimal allocations provided that $\partial b(x, y)/\partial y$ is a positive monotonic nonincreasing function of y for $x \in R$. In section 2 of this paper we show that for any Borel measurable local effectiveness function, incrementally optimal allocations are totally optimal whenever the target's probability distribution is given by a density function as in (1.1). In the case where the search space is countable, we prove that concavity of the local effectiveness function guarantees that incrementally optimal allocations are totally optimal. In addition, it is shown by counterexample that this property need not hold for countable search spaces if the local effectiveness function is not concave.

A search plan is called uniformly optimal if it maximizes the probability of detection at each instant during the search. In section 3, we show the existence of uniformly optimal search plans under additional hypotheses which are given there.

Our results hold in a more general situation than that of search theory. Thus, we introduce the following framework which is substantially the same as that in [6], one difference being that we deal only with Borel functions. Let R be a Borel subset of Euclidean n-space. We fix Borel functions L and U with $L \leq U$ which are defined on R. The functions L and U may take infinite values.

Define $\Omega = \{(x, y) : x \in R, |y| < \infty \text{ and } L(x) \leq y \leq U(x)\}$. We fix a real-valued Borel function e defined on Ω and the family Ξ of a.e. (with respect to Lebesgue measure) real-valued Borel functions q defined on R such that $L \leq q \leq U$. For $q \in \Xi$ we understand $e(\cdot, q(\cdot))$ to be a function from R to the reals. Define

$$\Phi = \Xi \cap \{q : e(\cdot, q(\cdot)) \text{ and } q \text{ are integrable}\},$$

and let

$$E(q) = \int_R e(x, q(x))dx \text{ and } C(q) = \int_R q(x)dx \qquad \text{for } q \in \Phi$$

All integration is Lebesgue integration. A $q^* \in \Phi$ is said to be *optimal* if

$$E(q^*) = \max\{E(q) : q \in \Phi \text{ and } C(q) = C(q^*)\}.$$

In the case where $L(x) = 0$, $U(x) = \infty$ for $x \in R$ and $e(x, y) = f(x)b(x, y)$ for $(x, y) \in \Omega$, $E(q)$ becomes the probability of detecting the target with allocation q and $C(q)$ becomes the amount of effort required by q. Then an optimal q^* maximizes the probability of detection which can be obtained with effort $C(q^*)$.

A function f defined on the real line is said to be *increasing* if $y \geq x$ implies $f(y) \geq f(x)$. A function f is said to be *concave* if for all x, y in the domain of f, $f(\alpha x + (1-\alpha)y)) \geq \alpha f(x) + (1-\alpha)f(y)$ for $0 \leq \alpha \leq 1$.

2. INCREMENTAL OPTIMIZATION

For $i = 1, 2, \ldots$, let $q_i \in \Phi$ be such that $q_1 \leq q_2 \leq \ldots$. Let $q_0 = L$. If

$$E(q_i) = \max \{E(q) : q \geq q_{i-1},\ q \in \Phi \text{ and } C(q) = C(q_i)\} \qquad \text{for } i = 1, 2, \ldots,$$

then we say that $(q_1, q_2, \ldots)$ is an *incrementally optimal* sequence. If q_i satisfies

$$E(q_i) = \max \{E(q) : q \in \Phi \text{ and } C(q) = C(q_i)\} \qquad \text{for } i = 1, 2, \ldots,$$

then $(q_1, q_2, \ldots)$ is said to be a *totally optimal* sequence. Define

$$\ell(x, y, \lambda) = e(x, y) - \lambda y \qquad \text{for } -\infty < \lambda < \infty, \text{ and } (x, y) \in \Omega.$$

The function ℓ is called a pointwise Lagrangian in [6] and λ is a Lagrange multiplier.

THEOREM 2.1: Let $(q_1, q_2, \ldots)$ be an incrementally optimal sequence such that for $i = 1, 2, \ldots$ $|E(q_i)| < \infty$ and $C(q_i)$ is in the interior of the range of C. Then $(q_1, q_2 \ldots)$ is a totally optimal sequence.

PROOF: By the definition of incremental optimality, q_1 is optimal. Thus, by Corollary 5.2 of [6], there exists a real number λ_1 such that for a.e. $x \in R$

$$\ell(x, q_1(x), \lambda_1) \geq \ell(x, y, \lambda_1) \qquad \text{for } |y| < \infty \text{ and } L(x) \leq y \leq U(x). \tag{2.1}$$

In other words a necessary condition for q_1 to be optimal is that it maximize a pointwise Lagrangian for some multiplier λ_1. Similarly, the incrementally optimal nature of q_2 implies the existence of a real number λ_2 such that for a.e. $x \in R$

$$\ell(x, q_2(x), \lambda_2) \geq \ell(x, y, \lambda_2) \qquad \text{for } |y| < \infty \text{ and } q_1(x) \leq y \leq U(x). \tag{2.2}$$

In order to prove that q_2 is optimal it is sufficient to find a real number λ such that for a.e. $x \in R$

$$\ell(x, q_2(x), \lambda) \geq \ell(x, y, \lambda) \qquad \text{for } |y| < \infty \text{ and } L(x) \leq y \leq U(x). \tag{2.3}$$

The sufficiency of (2.3) follows from a well known result concerning Lagrange multipliers (see, for example [3], [8] or Theorem 2.1 of [6]).

By (2.1) and (2.2)

$$\lambda_2(q_2(x) - q_1(x)) \leq e(x, q_2(x)) - e(x, q_1(x)) \leq \lambda_1(q_2(x) - q_1(x)) \qquad \text{for a.e. } x \in R. \tag{2.4}$$

Recall that $q_2 \geq q_1$. If $q_2(x) = q_1(x)$ for a.e. $x \in R$, then (2.3) holds for $\lambda = \lambda_1$. If $q_2(x) > q_1(x)$ for x in a set of positive measure, then (2.4) implies that $\lambda_2 \leq \lambda_1$. In this case for a.e. $x \in R$ and y such that $|y| < \infty$ and $L(x) \leq y \leq q_1(x)$, we have

$$0 \leq e(x, q_1(x)) - e(x, y) - \lambda_1(q_1(x) - y) \leq e(x, q_1(x)) - e(x, y) - \lambda_2(q_1(x) - y).$$

That is for a.e. $x \in R$,

$$\ell(x, y, \lambda_2) \leq \ell(x, q_1(x), \lambda_2) \leq \ell(x, q_2(x), \lambda_2) \qquad \text{for } |y| < \infty,\ L(x) \leq y \leq q_1(x). \tag{2.5}$$

Combining (2.5) and (2.2) we obtain (2.3) with $\lambda=\lambda_2$. Thus, q_2 is optimal. By repeating the argument for $q_3, q_4, \ldots$, the theorem is proved.

We now shift our attention to the case where R is a countable set. That is for some countable subset $\mathbf{J}$ of the integers, $R=\{x_j: j\in\mathbf{J}\}$. Let

$$\tilde{E}(q)=\sum_{j\in\mathbf{J}} e(x_j, q(x_j))$$

$$\tilde{C}(q)=\sum_{j\in\mathbf{J}} q(x_j).$$

Carry over the definitions of incrementally and totally optimal sequences in the obvious way. One may use the method of proof given in Theorem 2.1 to show that incrementally optimal sequences are totally optimal for the case where R is countable provided that the existence of a real number λ such that

$$\ell(x_j, q^*(x_j), \lambda)=\sup\{\ell(x_j, y, \lambda): |y|\leq\infty \text{ and } L(x)<y\leq U(x)\} \qquad \text{for } j\in\mathbf{J} \tag{2.6}$$

is a necessary condition for q^* to satisfy

$$\tilde{E}(q^*)=\max\{\tilde{E}(q): q\in\Phi \text{ and } \tilde{C}(q)=\tilde{C}(q^*)\}. \tag{2.7}$$

From Corollary 5.3 and Remark 2.3 of [6] we conclude that if $e(x_j, \cdot)$ is a concave function for $j\in\mathbf{J}$, then (2.6) is necessary for (2.7). Thus, we may state the following theorem.

THEOREM 2.2: If R is countable and $e(x_j, \cdot)$ is concave for $j\in\mathbf{J}$, then any incrementally optimal sequence is totally optimal.

The following example shows that one cannot remove the assumption that $e(x_j, \cdot)$ is concave in Theorem 2.2. The example also shows that (2.6) is not necessary for (2.7) when $e(x_j, \cdot)$ is not concave for $j\in\mathbf{J}$.

EXAMPLE 2.3: Let $R=\{1, 2\}$ be a doubleton set, $L=0$, $U(1)=2$, and $U(2)=\sqrt{3}$. Define

$$e(1, y)=\frac{1}{2}y \qquad 0\leq y\leq 2, \qquad e(2, y)=\begin{cases}\frac{1}{2}y & 0\leq y\leq 1\\ \frac{1}{4}y^2+\frac{1}{4} & 1<y\leq\sqrt{3}.\end{cases}$$

Note that both $e(1, \cdot)$ and $e(2, \cdot)$ are everywhere differentiable. For $0\leq T\leq 2+\sqrt{3}$, define

$$\eta(1, T)=\begin{cases}0, & 0\leq T\leq\sqrt{3}\\ T-\sqrt{3}, & \sqrt{3}\leq T\leq 2+\sqrt{3}\end{cases}$$

$$\eta(2, T)=\begin{cases}T, & 0\leq T\leq\sqrt{3}\\ \sqrt{3}, & \sqrt{3}<T\leq 2+\sqrt{3}.\end{cases}$$

Then $\eta(i, \cdot)$, $i=1, 2$, is increasing, and for each $T\geq 0$, $\tilde{C}(\eta(\cdot, T))=T$ and $\tilde{E}(\eta(\cdot, T))$ gives the maximum of $\tilde{E}(q)$ over all nonnegative functions q defined on $\{1, 2\}$ such that $\tilde{C}(q)=T$. Note that

for $q^* = \eta(\cdot, 1)$, (2.6) is *not* satisfied for any λ. An example of a function q^* which satisfies (2.7), but for which there is no λ satisfying (2.6) is also given in [8].

One may check that the sequence of allocations (q_1, q_2), where $q_1(1) = 1$, $q_1(2) = 0$ and $q_2(1) = 1$, $q_2(2) = 1$, is incrementally optimal. However, $\bar{E}(q_2) = 1$ and $\bar{C}(q_2) = 2$, while

$$\bar{E}(\eta(\cdot, 2)) = 2 - \tfrac{1}{2}\sqrt{3} > 1$$

so that q_2 is not optimal, i.e., (q_1, q_2) is not totally optimal.

In [2, p. 328] it is claimed that (in our notation) the existence of a function η defined on $R \times [0, S]$ such that $\eta(\cdot, T)$ is an optimal allocation of T amount of effort for each $0 \leq T \leq S$ and $\eta(x, \cdot)$ is increasing for $x \in R$ guarantees that incrementally optimal sequences are totally optimal. Example 2.3 shows that for discrete R this claim does not hold. If in addition to the existence of a function η satisfying the above conditions we have that for each amount of effort there is an almost everywhere *unique* optimal allocation of that effort, then any incrementally optimal sequence is totally optimal. Although not stated as such, this result is proven in [2].

Example 2.3 shows, of course, that optimal allocations need not be unique. Even when E and C are defined as integrals with respect to Lebesgue measure on n-space as is done in section 1, optimal allocations need not be unique. In fact, it is easy to see that if there exists a subset D of R having positive measure such that for $x \in D$ the graph of $e(x, \cdot)$ contains a nondegenerate straight-line segment of slope λ, then there are amounts of effort for which an optimal allocation of that effort is not almost everywhere unique.

REMARK 2.4: Let us return to the search situation described in section 1. That is, $L(x) = 0$, $U(x) = \infty$ for $x \in R$, $e(x, y) = f(x)b(x, y)$ for $(x, y) \in \Omega$. Suppose that an optimal allocation q_1 has been performed and that the search has failed to detect the target. Let f_1 be the posterior target location density given failure to detect the target. Thus

$$f_1(x) = \frac{f(x)\,[(1 - b(x, q_1(x))]}{1 - E(q_1)}. \tag{2.7}$$

For $x \in R$, let $b_1(x, \cdot)$ be the conditional local effectiveness function at x given that $q_1(x)$ search effort density was placed at x and the target not detected. Then

$$b_1(x, y) = \frac{b(x, q_1(x) + y) - b(x, q_1(x))}{1 - b(x, q_1(x))}. \tag{2.8}$$

Suppose that h is an allocation of effort which is added onto the original allocation q_1, so that the resulting total effort density is $q_1(x) + h(x)$ for $x \in R$. Then

$$E_1(h) = \int_R f_1(x)\, b_1(x, h(x))\, dx \tag{2.9}$$

is the conditional probability of detecting the target given that allocation q_1 failed.

Fix an increment of effort T. Suppose h^* has the property that $\int_R h^*(x)\, dx = T$ and

$$E_1(h^*) = \max \left\{ E_1(h) : h \geq 0 \text{ and } \int_R h(x)dx = T \right\}.$$

Then h^* is sometimes called a *conditionally optimal* search. If we let $q_2 = q_1 + h^*$, then we claim (q_1, q_2) is an incrementally optimal sequence. To see this, we observe that by (2.7) and (2.8),

$$E_1(h) = \frac{E(q_1 + h) - E(q_1)}{1 - E(q_1)}.$$

Thus maximizing E_1 subjected to $h \geq 0$ and $\int_R h(x)\,dx = T$ is equivalent to maximizing E subject to $q \geq q_1$ and $C(q) = C(q_1) + T$. The claim now follows from the definition of incremental optimality, and we see that the concepts of incremental and conditional optimality coincide for searches of the type discussed in this paper. Hence, under the conditions of Theorem 2.1 or 2.2, a sequence of conditionally optimal searches $(h_1, h_2, \ldots)$ produce, by setting $q_i = \sum_{k=1}^{i} h_k$ a totally optimal sequence $(q_1, q_2, \ldots)$ of search allocations.

3. EXISTENCE THEOREM

In this section we find conditions under which uniformly optimal search plans or allocation schedules exist. More precisely let J be an interval of real numbers. Then an *allocation schedule* over J is a Borel function η defined on $R \times J$ such that for $T \in J$, $\eta(\cdot, T) \in \Phi$ and for *a.e.* $x \in R$, $\eta(x, \cdot)$ is increasing. An allocation schedule η is *uniformly optimal* if

$$C(\eta(\cdot, T)) = T \text{ and } E(\eta(\cdot, T)) = \max\{E(q) : C(q) \leq T\} \qquad \text{for } T \in J. \tag{3.1}$$

This definition is a generalization of the definition of uniform optimality for search plans given by Arkin in [1]. In the special case where $E(q)$ gives the probability of detection resulting from the allocation of search effort, q, we call η a *search plan*. Then a uniformly optimal search plan maximizes the probability of detection at each instant during the search. In order to prove the existence of such allocation plans we define a notion of coverability similar to the one in [7].

Suppose p is a real-valued function defined on an interval J of real numbers. If p is concave, then throughout the interior of its domain, p' exists a.e. and is decreasing. Moreover, if p is continuous, then $p(t) - p(s) = \int_s^t p'(r)\,dr$ for $s, t \in J$. By an *extreme* point of a concave function p, we mean a point on its graph which does not lie on a chord joining two other points on the graph.

Define $m(x, \cdot)$ to be the minimal concave majorant of $e(x, \cdot)$ for all $x \in R$ for which such a majorant exists. We say that m covers e if the following conditions are satisfied.

(i) For a.e. $x \in R$, $m(x, \cdot)$ exists and is continuous.

(ii) m is a Borel function.

(iii) $e(x, y) = m(x, y)$ whenever $(y, m(x, y))$ is an extreme point of $m(x, \cdot)$.

Note that condition (iii) is equivalent to assuming that $e(x, \cdot)$ is upper semi-continuous at y such that $(y, m(x, y))$ is an extreme point of $m(x, \cdot)$. For $q \in \Phi$ we define

$$M(q) = \int_R m(x, q(x))\,dx$$

whenever the integral on the right exists.

Differentiation is always with respect to the last component of the argument, and is denoted by a prime, e.g., for $(x, y)\epsilon\Omega$, $e'(x, y) = \lim_{h\to 0} [e(x, y+h) - e(x, y)]/h$. Let m cover e. If a function $q\epsilon\Phi$ and a real number λ satisfy, for a. e. $x\epsilon R$,

$$m'(x, y) \geq \lambda \qquad \text{for a. e. } y \text{ such that } L(x) < y < q(x)$$

$$m'(x, y) \leq \lambda \qquad \text{for a. e. } y \text{ such that } q(x) < y < U(x),$$

then we say that the pair (q, λ) satisfies the Neyman-Pearson inequalities. When $e(x,\cdot)$ and $m(x,\cdot)$ are increasing and $U(x) = \infty$, it is convenient to define

$$e(x, \infty) = \lim_{y\to\infty} e(x, y) \text{ and } m(x, \infty) = \lim_{y\to\infty} m(x, \infty).$$

Before proceeding with our main existence result, we prove two lemmas which will be useful in this section. Lemma 3.1 relates closely to Theorem 1 and Remark 3 of [5].

LEMMA 3.1: Let m cover e. If there is a $q^*\epsilon\Phi$ such that $E(q^*) > -\infty$ and $\lambda \geq 0$ such that for a. e. $x\epsilon R$

$$(3.2)\quad \begin{cases} \text{(i)} \quad m'(x, y) \geq \lambda & \text{for a. e. } y \text{ such that } L(x) < y < q^*(x) \\ \text{(ii)} \quad m'(x, y) \leq \lambda & \text{for a. e. } y \text{ such that } q^*(x) < y < U(x) \\ \text{(iii)} \quad e(x, q^*(x)) = m(x, q^*(x)), & \end{cases}$$

then

$$(3.3)\qquad E(q^*) = \max \{E(q) : C(q) \leq C(q^*)\}.$$

PROOF: By (3.2) (iii), $M(q^*)$ exists. It is an easily shown Neyman-Pearson result (see Theorem 1 of [5]) that for $\lambda \geq 0$, (i) and (ii) imply

$$(3.4)\qquad M(q^*) = \max \{M(q) : C(q) \leq C(q^*)\}.$$

Suppose that there is an $r\epsilon\Phi$ such that $E(r) > E(q^*)$ and $C(r) \leq C(q^*)$. Since m majorizes e, we have

$$M(r) \geq E(r) > E(q^*) = M(q^*),$$

which contradicts (3.4). This proves the lemma.

For $\lambda > 0$ and x such that $m(x,\cdot)$ exists, we define

$$\varphi_u(x, \lambda) = \sup \{y : y = L(x) \text{ or } m'(x, y) \geq \lambda\}$$

$$\varphi_\ell(x, \lambda) = \inf \{y : y = U(x) \text{ or } m'(x, y) \leq \lambda\}.$$

Then for $\lambda > 0$, we let

$$I_\ell(\lambda) = \int_R \varphi_\ell(x, \lambda)dx, \qquad I_u(\lambda) = \int_R \varphi_u(x, \lambda)dx.$$

The functions φ_ℓ and φ_u will be our main tools for constructing solutions to the constrained extremal problems considered here. The following lemma displays some of the properties of these functions.

LEMMA 3.2: Suppose m covers e and for a.e. $x\epsilon R$, $e(x,\cdot)$ is increasing. If $-\infty < E(L) \leq E(U) < \infty$ and $|C(L)| < \infty$, then the following hold:

(a) $\varphi_u(\cdot, \lambda)\epsilon\Phi$ and $\varphi_\ell(\cdot, \lambda)\epsilon\Phi$ for $\lambda > 0$.

(b) I_ℓ and I_u are finite and decreasing.

(c) $\varphi_\ell(x, \cdot)$ and I_ℓ are right continuous and $\varphi_u(x, \cdot)$ and I_u are left continuous.

(d) $(\varphi_\ell(\cdot, \lambda), \lambda)$ and $(\varphi_u(\cdot, \lambda), \lambda)$ satisfy the Neyman-Pearson conditions.

(e) A pair (q, λ), where $q\epsilon\Phi$ and $\lambda \geq 0$, satisfies the Neyman-Pearson inequalities if, and only if,

$$\varphi_\ell(x, \lambda) \leq q(x) \leq \varphi_u(x, \lambda) \qquad \text{for a.e. } x\epsilon R.$$

(f) For any $\lambda > 0$, we may find a Borel function α defined on $R \times [I_\ell(\lambda), I_u(\lambda)]$, such that

(1) $\alpha(x, \cdot)$ is increasing for a.e. $x\epsilon R$,

(2) $C(\alpha(\cdot, T)) = T$ for $I_\ell(\lambda) \leq T \leq I_u(\lambda)$,

(3) $(\alpha(\cdot, T), \lambda)$ satisfies the Neyman-Pearson inequalities for all $I_\ell(\lambda) \leq T \leq I_u(\lambda)$.

(g) $\lim_{\lambda \to \infty} I_u(\lambda) = C(L)$.

(h) For $\lambda > 0$ and x such that $m(x, \cdot)$ exists, $(\varphi_u(x, \lambda), m(x, \varphi_u(x, \lambda)))$, and $(\varphi_\ell(x, \lambda), m(x, \varphi_\ell(x, \lambda)))$ are extreme points of $m(x, \cdot)$.

PROOF: A straightforward verification shows that $\varphi_\ell(\cdot, \lambda)$ and $\varphi_u(\cdot, \lambda)$ are Borel functions for each $\lambda > 0$ and that (a) holds. Thus, the integrals, $I_\ell(\lambda)$ and $I_u(\lambda)$ are well defined for each $\lambda > 0$.

For a.e. $x\epsilon R$, the following hold. Since $e(x, \cdot)$ is increasing, $m(x, \cdot)$ is increasing. If $U(x)$ is finite, then $(U(x), m(x, U(x)))$ is an extreme point and $m(x, U(x)) = e(x, U(x))$. If $U(x) = \infty$, then the increasing nature of $e(x, \cdot)$ and the minimal nature of $m(x, \cdot)$ yields $m(x, \infty) = e(x, \infty)$. Since $|C(L)| < \infty$, $L(x)$ is finite and $m(x, L(x)) = e(x, L(x)\cdot)$.

To prove (b), we observe that

$$-\infty < E(L) = M(L) \leq E(U) = M(U) < \infty.$$

Thus, for a.e. $x\epsilon R$, $m(x, L(x))$ and $m(x, U(x))$ are finite.

Since $m(x, \cdot)$ is increasing, we have for a. e. $x\epsilon R$,

$$m(x, U(x)) - m(x, L(x)) \geq \int_{L(x)}^{z} m'(x, y)dy \geq (z - L(x))\, m'(x, z) \qquad \text{for } L(x) < z < U(x).$$

Thus, $m'(x, z) \leq (m(x, U(x)) - m(x, L(x)))/(z - L(x))$, and it follows that

$$\varphi_u(x, \lambda) \leq \frac{1}{\lambda}[m(x, U(x)) - m(x, L(x))] + L(x) \qquad \text{for } \lambda > 0.$$

Hence,

$$-\infty < C(L) \leq I_\ell(\lambda) \leq I_u(\lambda) \leq \frac{1}{\lambda}[M(U) - M(L)] + C(L) < \infty \qquad \text{for } \lambda > 0$$

which proves that I_ℓ and I_u are finite. The decreasing nature of $m'(x, \cdot)$ for a. e. $x \in R$ guarantees that $\varphi_u(x, \cdot)$ and $\varphi_\ell(x, \cdot)$ are decreasing for a. e. $x \in R$. Thus, (b) follows.

The left continuity of $\varphi_u(x, \cdot)$ and the right continuity of $\varphi_\ell(x, \cdot)$ for a. e. $x \in R$ follow from their definitions and the decreasing nature of $m'(x, \cdot)$. The monotone convergence theorem and the finiteness of I_u and I_ℓ may be used to show the left and right continuity of I_u and I_ℓ, respectively. Thus, (c) holds.

Properties (d) and (e) follow directly from the definition of φ_ℓ and φ_u. In order to prove (f), we use a device of Arkin's [1] and define for $0 \leq s \leq \infty$

$$h_\lambda(x, s) = \begin{cases} \varphi_u(x, \lambda) & \text{if } |x| < s \\ \varphi_\ell(x, \lambda) & \text{if } |x| \geq s \end{cases}$$

and

$$H_\lambda(s) = \int_R h_\lambda(x, s)dx.$$

By the monotone convergence theorem, H_λ is continuous. Moreover, H_λ is increasing and

$$H_\lambda(0) = I_\ell(\lambda), \qquad H_\lambda(\infty) = I_u(\lambda).$$

Thus for $I_\ell(\lambda) \leq T \leq I_u(\lambda)$, we may choose $\xi(T)$ such that $H_\lambda(\xi(T)) = T$. Defining $\alpha(x, T) = h_\lambda(x, \xi(T))$ for $x \in R$, we see that α satisfies conditions (1) and (2) of (f). Condition (3) follows from (e). Property (g) follows easily from the monotone convergence theorem and the definition of φ_u. Property (h) may be verified from the definitions of φ_ℓ, φ_u, and an extreme point. This completes the proof.

THEOREM 3.3: Suppose m covers e, and for a. e. $x \in R$, $e(x, \cdot)$ is increasing. If $-\infty < E(L) \leq E(U) < \infty$ and $|C(L)| < \infty$, then there exists a uniformly optimal allocation schedule η over $[C(L), C(U))$.

PROOF: We consider first the case where $I_u(0) \equiv \lim_{\lambda \to 0} I_u(\lambda) = C(U)$. The case $I_u(0) < C(U)$ requires only routine modifications which are discussed at the end of the proof. We take $\eta(\cdot, C(L)) = L$.

Since I_u is monotone, it has only a countable number of discontinuities. Let K be a countable index set such that $\{\lambda_k : k \in K\}$ is the set of discontinuity points of I_u. Let $J_k = [I_\ell(\lambda_k), I_u(\lambda_k]$ for $k \in K$. The intervals J_k are disjoint and are the jump intervals at the discontinuity points of I_u. For $T \in (C(L), C(U)) - \bigcup_{k \in K} J_k$, let

$$\lambda^*(T) = \sup\{\lambda : I_u(\lambda) = T\}.$$

By the left continuity of I_u, $I_u(\lambda^*(T)) = T$.

For $T \in J_k$, let $\lambda^*(T) = \lambda_k$ and choose a function α_k defined on $R \times J_k$ to have the properties of α in (f) of Lemma 3.2. Define

$$\eta(x, T) = \begin{cases} \varphi_u(x, \lambda^*(T)) & \text{if } C(L) < T < C(U) \text{ and } T \notin \bigcup_{k \in K} J_k \\ \alpha_k(x, T) & \text{if } T \in J_k \text{ and } k \in K. \end{cases}$$

Then for each $C(L) < T < C(U)$, $C(\eta(\cdot, T)) = T$ and $(\eta(\cdot, T), \lambda^*(T))$ satisfies the Neyman-Pearson conditions. Since m covers e and property (h) of Lemma 3.2 holds, we have that for each $C(L) < T < C(U)$, $e(x, \eta(x, T)) = m(x, \eta(x, T))$ for a.e. $x \epsilon R$. Thus, by Lemma 3.1, η satisfies 3.1.

To verify that $\eta(x, \cdot)$ is increasing for a.e. $x \epsilon R$, we let R' be the set of $x \epsilon R$ such that $m(x, \cdot)$ exists. Then by the fact that m covers e, $R - R'$ has measure 0. Suppose it is not the case that $\eta(x, \cdot)$ is increasing for a.e. $x \epsilon R$. Then there is an $x \epsilon R'$ and numbers S and T, such that $C(L) < T < S < C(U)$ and

$$\eta(x, S) < \eta(x, T). \tag{3.4}$$

Since $(\eta(\cdot, T), \lambda^*(T))$ and $(\eta(\cdot, S), \lambda^*(T))$ satisfy the Neyman-Pearson inequalities for all $x \epsilon R'$, we have

$$\lambda^*(T) \leq m'(x, y) \leq \lambda^*(S) \qquad \text{for a.e. } y \text{ such that } \eta(x, S) < y < \eta(x, T).$$

One may check that λ^* is a decreasing function, so that $\lambda^*(T) = \lambda^*(S)$. Thus, for some $k \epsilon K$, T and S are both in the closure of J_k. However, $\eta(x, \cdot)$ is constructed by property (f) of Lemma 3.2 to be increasing on the closure of J_k. This contradicts (3.4) and proves the theorem for the case where $I_u(0) = C(U)$.

If $I_u(0) < C(U)$, we proceed as before for $C(L) < T \leq I_u(0)$. We then define

$$\varphi_u(x, 0) = \lim_{\lambda \to 0} \varphi_u(x, \lambda).$$

From the increasing nature of $e(x, \cdot)$, it follows that if $q \epsilon \Phi$ and if $q(x) \geq \varphi_u(x, 0)$ for $x \epsilon R$, then q will satisfy (3.2) with $\lambda = 0$. Hence, to complete the definition of $\eta(x, \cdot)$ for $I_u(0) < T < C(U)$, one need only choose η so that $\eta(x, T) \geq \varphi_u(x, 0)$ and $C(\eta(\cdot, T)) = T$ which may be easily done. This completes the proof.

Observe that the hypotheses of Theorem 3.3 may be weakened to require that $m(x, \cdot)$ rather than $e(x, \cdot)$ be increasing for a.e. $x \epsilon R$. The theorem remains unchanged except that η must be restricted to $[C(L), I_u(0)]$. This is no real restriction since for $q \geq \varphi_u(\cdot, 0)$, $E(q) \leq E(\varphi_u(\cdot, 0))$.

Theorem 2 of Arkin [1] is similar to Theorem 3.3 above with the exception that [1] claims that there exists a function β such that

$$\eta(x, T) = \int_0^T \beta(x, s) ds, \qquad C(\eta(\cdot, T)) = T,$$

and η is uniformly optimal. However, the following is a counterexample to Theorem 2 of [1]. (Moreover, the proof in [1] is not sufficient to show the truth of Theorem 3.3.)

Let $R = [0, 1]$, $L = 0$, $U = \infty$, and

$$e(x, y) = \begin{cases} 0, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}$$

for $x \epsilon R$. It is clear that any uniformly optimal search plan η must have the property that for a.e. $x \epsilon R$, $\eta(x, \cdot)$ jumps from 0 to 1 at some point T, but there is no function β which produces this behavior for η.

Under the conditions of Theorem 3.3. we have shown that there exists, for any $C(L) \leq T < C(U)$, a q^* such that $C(q^*) = T$ and $E(q^*) = \max\{E(q) : q\epsilon\Phi \text{ and } C(q) \leq T\}$. Theorem 8 of [7] provides a similar existence result whenever m covers e and $-\infty < C(L) \leq C(U) < \infty$. In comparison, Theorem 3.3 of this paper removes the restriction that $C(U) < \infty$, but adds monotonicity conditions on $e(x, \cdot)$ and boundedness conditions on E. In [6] there is also a discussion of related existence theorems.

One might conjecture that Theorem 3.3 would remain true without assuming that $e(x, \cdot)$ is increasing, provided that we assumed $|E(q)| < B$ for some number B and all $q\epsilon\Phi$. Similarly, one might conjecture that the restriction $C(L) > -\infty$ could be omitted. However, the following two counterexamples show both of these conjectures to be false.

EXAMPLE 3.4: Let $R = [1, \infty)$, $L = 0$, and $U(x) = x + 1/x^2$ for $x\epsilon R$.
For $x\epsilon R$, define

$$e(x, y) = \begin{cases} y, & 0 \leq y \leq 1/x^2 \\ \dfrac{1}{x^2} - \dfrac{y - 1/x^2}{x^3}, & 1/x^2 < y \leq x + 1/x^2. \end{cases}$$

Note that $m = e$ and that $|E(q)| \leq 1$ for all $q\epsilon\Phi$. Suppose q^* is optimal and

$$\infty > C(q^*) > 1.$$

By Corollary 7.2 of [6] there exists a λ such that

$$e(x, q^*(x)) - \lambda q^*(x) = \sup\{e(x, y) - \lambda y : 0 \leq y < x + 1/x^2\} \qquad \text{for a.e. } x\epsilon R. \tag{3.5}$$

Since $e(x, \cdot)$ is concave for $x\epsilon R$, this implies

$$e'(x, y) \geq \lambda \qquad \text{for } 0 < y < q^*(x)$$

$$e'(x, y) \leq \lambda \qquad \text{for } q^*(x) < y < x + 1/x^2, \qquad \text{for a.e. } x\epsilon R.$$

One may check that if $\lambda \geq 0$, $C(q^*) \leq 1$. Thus, the above λ must be negative. It follows from the above inequalities that $q^*(x) = x + 1/x^2$ for $x^3 > -1/\lambda$. Thus, $C(q^*) = \infty$ which contradicts our assumption that $C(q^*) < \infty$. Thus, one cannot replace the monotonicity of $e(x, \cdot)$ by boundedness of E in Theorem 3.3.

EXAMPLE 3.5: Let $R = [1, \infty)$, $L = -1$, $U = 1$, and

$$e(x, y) = y/x^2 \qquad \text{for } -1 \leq y \leq 1.$$

Observe that $e = m$ and all the conditions of Theorem 3.3 are satisfied except that $C(L) = -\infty$. Suppose q^* is optimal and $C(q^*)$ is finite. Again by Corollary 5.2 of [6] there exists a λ such that

$$e(x, q^*(x)) - \lambda q^*(x) = \sup\{e(x, y) - \lambda y : -1 \leq y \leq 1\} \qquad \text{for a.e. } x\epsilon R.$$

Hence,

$$e'(x, y) \geq \lambda, \qquad -1 < y < q^*(x)$$

$$e'(x, y) \leq \lambda, \qquad q^*(x) < y < 1, \qquad \text{for a.e. } x \epsilon R.$$

It follows that

$$q^*(x) = \begin{cases} 1 & \text{for } x^2 < 1/\lambda \\ -1 & \text{for } x^2 > 1/\lambda. \end{cases}$$

Hence, either $C(q^*) = -\infty$ or $C(q^*) = +\infty$ contrary to the assumption that $C(q)$ is finite. Thus, we cannot omit the condition $C(L) > -\infty$ in Theorem 3.3.

REMARK 3.6: In the search theory case where $L(x) = 0$, $U(x) = \infty$ for $x \epsilon R$ and $e(x, y) = f(x)\, b(x, y)$ for $(x, y) \epsilon \Omega$, the conditions of Theorem 3.3 will be satisfied if $b(x, \cdot)$ is right continuous for $x \epsilon R$. Since $b(x, \cdot)$ is increasing and $|E(q)| \leq 1$, for $q \epsilon \Phi$, the only condition that is not obviously satisfied is the coverability condition. However, since $b(x, \cdot)$ is increasing and right-continuous, it is upper semi-continuous. Thus, $e(x, \cdot)$ has a minimal concave majorant $m(x, \cdot)$ which is continuous, and one may check that $e(x, y) = m(x, y)$ whenever $(y, m(x, y))$ is an extreme point of $m(x, \cdot)$. It can be shown that since e is Borel, m is a.e. equal to a Borel function. Thus the conditions for coverability are satisfied. It follows that whenever the local effectiveness function is right continuous and the target location distribution is given by a density function on Euclidean n-space, a uniformly optimal search plan exists. Note that uniformly optimal search plans may be used to produce sequences which are both incrementally and totally optimal.

REFERENCES

[1] Arkin, V. I., "Uniformly Optimal Strategies in Search Problems," Theor. Probability Appl. *2*, 674–680 (1964).

[2] Dobbie, James M., "Search Theory: A Sequential Approach," Nav. Res. Log. Quart. *4*, 323–334 (Dec. 1963).

[3] Everett, Hugh, "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," Operations Res. *11*, 399–417 (1963).

[4] Koopman, B. O., "The Theory of Search: III. The Optimum Distribution of Searching Effort," Operations Res. *5*, 613–626 (1957).

[5] Wagner, D. H., "Non-Linear Functional Versions of the Neyman-Pearson Lemma," SIAM Rev. *1*, 52–65 (June 1969).

[6] Wagner, D. H. and L. D. Stone, "Necessity and Existence Results on Constrained Optimization of Separable Functionals by a Multiplier Rule," To appear in SIAM J. Control, 12, (1974).

[7] Wagner, D. H. and L. D. Stone, "Optimization of Allocations Under a Coverability Condition," To appear, SIAM J. Control, 12, (1974).

[8] Zahl, S., "An Allocation Problem With Applications to Operations Research and Statistics," Operations Res. *11*, 426–441 (1963).

AN APPROACH TO THE ALLOCATION OF COMMON COSTS OF MULTI-MISSION SYSTEMS*

Robert Thomas Crow

School of Management
State University of New York at Buffalo

ABSTRACT

Many Naval systems, as well as other military and civilian systems, generate multiple missions. An outstanding problem in cost analysis is how to allocate the costs of such missions so that their true costs can be determined and resource allocation optimized. This paper presents a simple approach to handling this problem for single systems. The approach is based on the theory of peak-load pricing as developed by Marcel Boiteux. The basic principle is that the long-run marginal cost of a mission must be equal to its "price." The implication of this is that if missions can cover their own marginal costs, they should also be allocated some of the marginal common costs. The proportion of costs to be allocated is shown to a function of not only the mission-specific marginal costs and the common marginal costs, but also of the "mission price." Thus, it is shown that measures of effectiveness must be developed for rational cost allocation. The measurement of effectiveness has long been an intractable problem, however. Therefore, several possible means of getting around this problem are presented in the development of the concept of relative mission prices.

THE PROBLEM

This paper is an attempt to provide a new method of allocating the common costs of new investments in a multi-mission system to individual missions in a way that is (1) operational, (2) objective, and (3) defensible from the point of view of efficient resource allocation.† The most important reason for allocating common costs to individual missions is to provide guidance for procurement of systems and to estimate the costs of accomplishing given missions by alternative systems. If common costs are properly allocated, it is possible to estimate the true costs of accomplishing a mission with one system compared to another.

For an illustration of the problem, consider Table 1. Five systems are shown that can be combined to accomplish three missions. Systems A, B, and C are single-mission systems and therefore, by definition, have no common costs. Systems D and E are multi-mission systems characterized by significant common costs, as well as incremental costs which are specific to each mission. How much does it cost to accomplish the missions by the use of multi-mission systems? Which systems—single-mission, multi-mission, or some combination—should be procured? Obviously, military systems problems are too complex to be accurately characterized by such simple questions. Yet it appears that in day-to-

*The work on which this article is based was performed for the Chief of Naval Operations, Systems Analysis, as part of Contract No. N00014–70–C–0086 with Mathematica, Inc. This paper is a revision of portions of a report [4] prepared for that contract.

† Common costs are defined as those which are incurred by a single system regardless of which of a number of missions is being performed. They may arise from either operation or investment. It should be clear that, since the subject is *investment decisions*, it is the *incremental* common costs that are to be allocated.

day operations, they are often posed in this fashion—for first-approximation purposes, if not for final decisions.

TABLE 1. *Comparison of Costs of Single- and Multi-Mission Systems*

Costs	Systems				
	A	B	C	D	E
Common	—	—	—	50	60
Mission 1	65	—	—	20	—
Mission 2	—	45	—	30	30
Mission 3	—	—	60	—	10
Totals	65	45	60	100	100

As an example of the consequences of misallocating common costs, if in a given system all common costs are allocated to a particular important mission, it may appear that the system's capability in that mission is costly relative to other systems and it may not be purchased. Even if it is purchased, it may be underutilized. On the other hand, other mission capabilities of the system may appear to be less costly than they are in fact, leading to overpurchase or overutilization.

In general, the allocation of common costs has been avoided by economists. Most work on common costs has focused on short-run marginal cost analysis and hence only on the costs that are variable for each specific output (or mission, in our context).* This focus on short-run problems side-steps the issue of how to handle common costs in investment decisions, where there are normally several alternative courses of action and no costs are fixed. For pricing and other types of decisions, in both the private and public sectors, if other than short run marginal or incremental prices are considered, reliance has usually been placed on arbitrary allocations of common costs to some or all outputs. This was the general rule until 1949, when Marcel Boiteux wrote an ingenious and basic article on peak-load pricing for electrical utilities, translated into English in 1964 (Ref. [2]). In the approach presented here, two principles will be emphasized: (1) (following Boiteux) efficient allocation of common costs can and must be based on "marginal" conditions; and (2) efficient allocation must be based explicitly on considerations of how well a given system performs a given mission, as well as an assessment of the need for capability in that mission.

SOME BASIC PRINCIPLES

A basic theorem of resource allocation is that in a competitive economy the maximation of output in equilibrium occurs when price (P) equals the cost of producing one additional unit of output (marginal cost or MC). The $P = MC$ output can be proven to be optimal if competitive conditions hold throughout the economy (assuming the existence of U-shaped average cost curves). This is a fundamental principle in the ensuing discussion on cost allocation.

*For an example, see Ref. [3, ch. 5]. Professor William Baumol has pointed out in correspondence that there have been a number of recent exceptions to this general rule for civilian systems, such as utilities. In particular, see Ref. [6].

Furthermore, it is necessary to recognize that *before any costs are sunk, they are all variable.* Thus, in planning for investment, the criterion that price must equal marginal cost means that the marginal cost measure *must include investment cost.* Assuming for the moment that a set of "prices" can be established for the value of accomplishing a military mission, the widely used criterion of price equal to short-run marginal costs (marginal operating costs) is only valid as the *minimum* price at which a particular system should be used in a given mission. In other words, for a given system, if the mission is not needed sufficiently to be worth sacrificing enough resources to pay for its marginal operating costs, then the mission should not be performed.

An example of the possible consequences of using only short-run marginal costs in system decisions is shown in Figures 1 and 2. The amount of output of each two systems is q_o and the vertical distances Of_1 and Of_2 represent investment costs, while the vertical distance of the shaded areas represent marginal costs. In this case, it is clear that short-run marginal costs are lower for system 2 than system 1; however, in the long run, the total cost of system 1 is less because its investment costs are lower. Therefore, it should be chosen in spite of its higher short-run marginal costs.

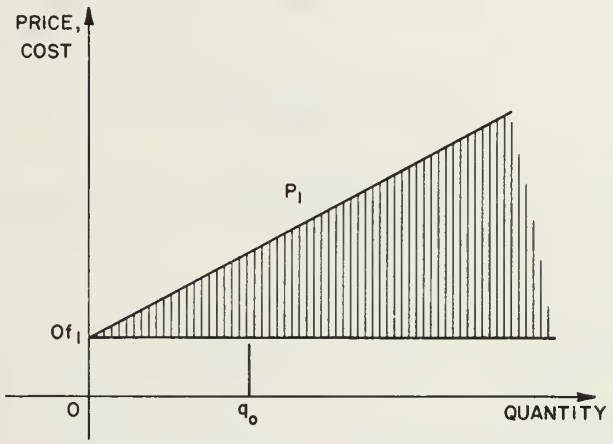

FIGURE 1. Output with high marginal operating costs and low marginal investment costs, System 1.

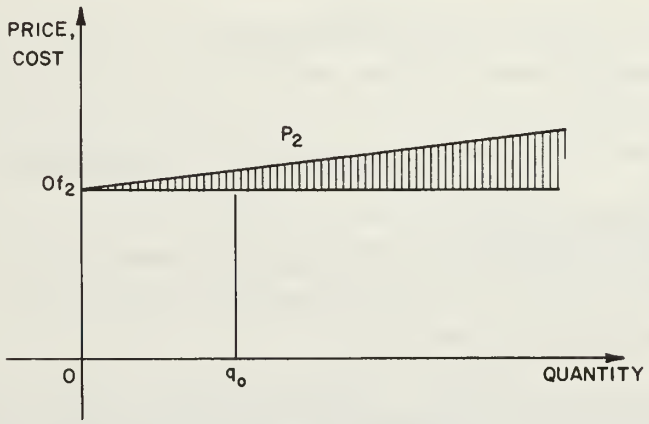

FIGURE 2. Output with low marginal operating costs and high marginal investment costs, System 2.

In order to include investment costs in the marginal cost measures, it is necessary to distinguish long-run from short-run costs. Long-run costs represent situations where output is variable through investment, rather than by changing utilization of existing capacity (which can occur over the short-run). The question is how to reconcile the short-run condition of price equal to marginal cost with the necessary condition of long-run marginal costs being covered. The solution is deceptively simple: for a given objective to be achieved, purchase that number of units of the system at which the price of each unit equals *both* the short-run and long-run marginal costs. It is necessary to demonstrate this assertion.*

Real systems often are relatively inflexible in their capability. That is, after a point, significant increases in operating costs will expand output very little, i.e., the cost curves have an "elbow" where they become vertical when capacity is reached. It can be proven that for a family of such short-run curves, representing either expanded units or additional units, the envelope of the short-run curves is the long-run total-cost curve. This is illustrated in Figure 3. This is the case that will be dealt with

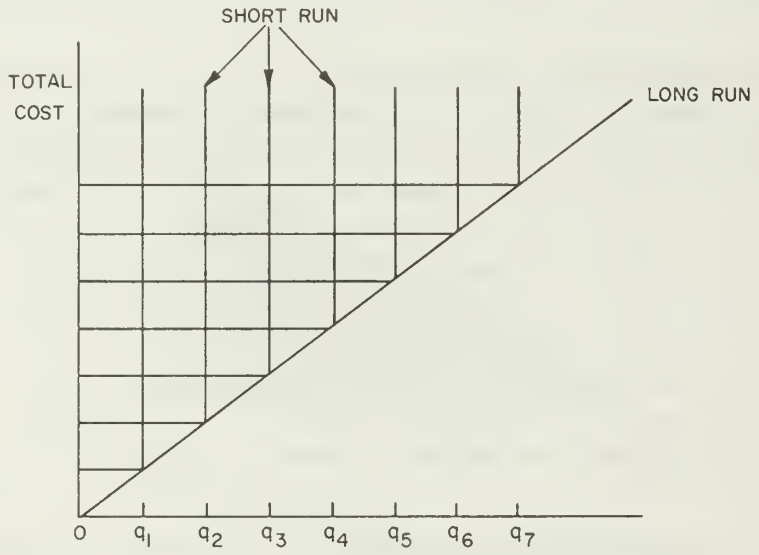

FIGURE 3. Relationship of long-run to short-run total costs.

here in demonstrating that optimum resource allocation calls for equality of short- and long-run marginal costs.† In the ensuing discussion it will be useful to identify short-run costs with operating costs, and long-run costs as operating *plus* investment costs. In dealing with a system with very little flexibility, we are able to illustrate a means of approximating the solution for the case in which the system is completely inflexible.

The following notation will be used:

C: total cost variable over the short run (only operating costs)

C_o: total cost variable over the long run (operating plus investment costs)

C_f: total investment cost (fixed in the short run)

*A rigorous treatment is presented in an appendix to Ref. [2].

†This can be relaxed, but at the sacrifice of simplicity in exposition and application. See appendix to Boiteux [2].

q: output of a system
q_o: capacity which meets some given requirement
z: marginal operating cost (variable in the short run) $= dC/dq$
x: marginal investment cost (variable in the long run) $= dC_f/dq_o$
y: marginal long run cost $= dC_o/dq_o$
w: an approximation of z for $q < q_o$, defined by $(C - C_f)/q$, i.e., the average *avoidable* cost.

For $q < q_o$, let the total cost function be*

$$C = C_f + wq. \tag{1}$$

Over the long run, adjustment can be made to the required capacity (i.e., investment can be varied). Therefore (1) can be expressed as

$$C_o = C_f(q_o) + wq_o, \tag{2}$$

and, assuming that w does not vary with the number of units,

$$\frac{dC_o}{dq_o} = \frac{dC_f}{dq_o}(q_o) + w = y. \tag{3}$$

That is, the long-run marginal cost of a number of units is

$$y = x + w, \tag{4}$$

by the definitions of y, x, and w, i.e., the long-run marginal costs are equal to the sum of the marginal investment cost and the average avoidable cost.

If there is true inflexibility at q_o, short-run marginal cost is indeterminate. However, if there is a slight bit of flexibility, z becomes very large as $q \to q_o$. Therefore, it becomes equal to y at some point on its vertical arm. At this point, long-run and short-run marginal are equal, which establishes the solution of price equal to long-run and short-run marginal costs,

$$p = y = x + w = z. \tag{5}$$

This is illustrated in Figure 4.

The Basic Principle of Common Cost Allocation

Assume that a given system can perform two missions, and its investment cost is common to both. That is, there are no characteristics of the system that can be attributed uniquely to one mission or another.†

---

* This section follows Boiteux, [2] closely. Any errors are likely to be mine, not his.

† Linearity is assumed for convenience and because in applied work one often has only linear approximations in any case. The linearity assumption does not appear to have a material effect on the analysis.

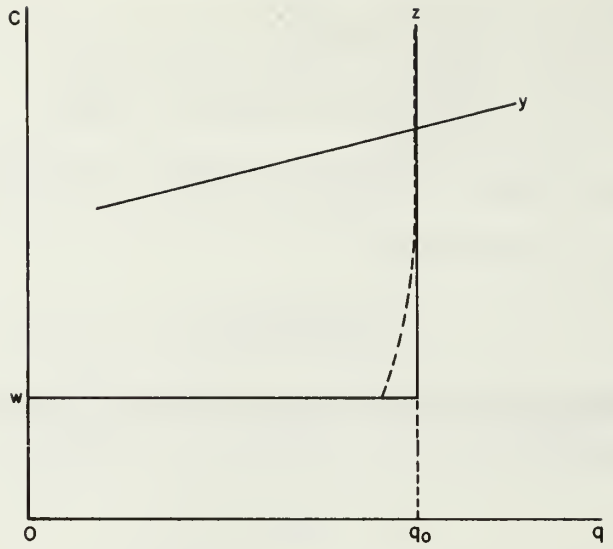

FIGURE 4. Equality of short-run and long-run marginal costs at capacity output.

In the case of two missions being performed by a single system, the cost functions for each of them, with flexible capacity, are

(6a) $$C_1 = C_f(q_o) + w_1 q_1$$

and

(6b) $$C_2 = C_f(q_o) + w_2 q_2,$$

where q_1 and q_2 represent the output of missions 1 and 2, respectively. Given a particular output requirement, say q_1, the requirement for optimum resource allocation is that the marginal long-run cost for both missions together be minimized, i.e.,

(7) $$\frac{\partial(C_1 + C_2)}{\partial q_o} = 0.$$

To establish the equality of long- and short-run marginal costs, consider the differential of $C_1 + C_2$,

(8) $$d(C_1 + C_2) = \frac{\partial C_1}{\partial q_1} dq_1 + \frac{\partial C_2}{\partial q_2} dq_2 + 2\frac{\partial(C_1 + C_2)}{\partial q_o} dq_o.$$

The first two terms on the right hand side are short-run marginal costs, z_1 and z_2, times their respective variations in output. The third term must be equal to zero for cost minimization to hold.

The differential $d(C_1 + C_2)$ may also be written as

(9) $$dC_1 + dC_2 = \left[\frac{\partial C_f}{\partial q_o} dq_o + w_1 dq_1\right] + \left[\frac{\partial C_f}{\partial q_o} dq_o + w_2 dq_2\right]$$

where $dq_1 = dq_2 = dq_o$,* (9) may be divided by dq_o to yield

$$\frac{dC_1 + dC_2}{dq_o} = 2\frac{\partial C_f}{\partial q_o} + w_1 + w_2. \tag{10}$$

Since $\partial C_f/\partial q_o$ has been defined above as x, the optimum condition of equality of short- and long-run marginal costs is

$$z_1 + z_2 = 2x + w_1 + w_2. \tag{11}$$

For the prices of two missions to be equal to the long-run marginal costs of a given system capacity,

$$p_1 + p_2 = 2x + w_1 + w_2. \tag{12}$$

The allocation of common investment costs follows immediately from (12) which implies,

$$(p_1 - w_1)/2x + (p_2 - w_2)/2x = 1. \tag{13}$$

That is, the share of common costs to be allocated to each mission performed by the system is equal to the corresponding term in (13).

Before turning to the problem of how the prices of missions are to be determined, it is instructive to consider two particular cases. One is the case where one of the missions (mission 2, for example) has a price such that it is just worth its marginal operating costs, $p_2 = w$. In this case all common costs are allocated to mission 1 by (13). This is illustrated in Figure 5.

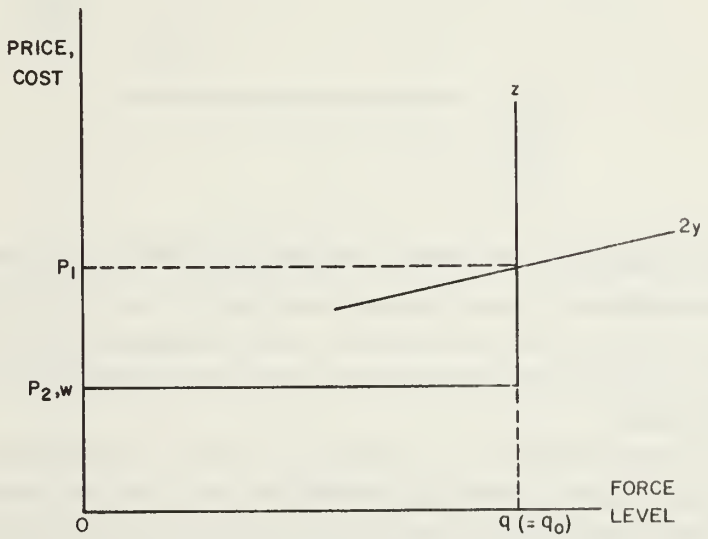

FIGURE 5. One system bearing all common costs and the other system bearing only operating costs.

*Boiteux [2, appendix] has shown that this is likely to be a reasonably good approximation over a wide range of conditions.

It is important to note that this is a generalization of the "two-ship" method of allocating investment costs currently in use (Grey, [5] pp. 2, 3, sec. 2–4). In this method, "major" missions are allocated the entire common investment cost of the system plus their incremental investment costs, while "minor" missions are allocated only their incremental investment costs.

In the case shown in Figure 6, the problem is different in that each mission's price is such as to be able to cover at least a portion of its common marginal investment costs. In this case, the question arises: Is one mission allocated only operating costs and the other allocated both its operating and the common investment costs? And further, if the common investment costs are shared, how are they divided? The answer is that since p_1 and p_2 both exceed w (assumed to be the same for both missions to keep the diagram simple), common costs must be allocated according to (13), where both terms will be greater than zero. Thus each mission will bear a share of the common cost.

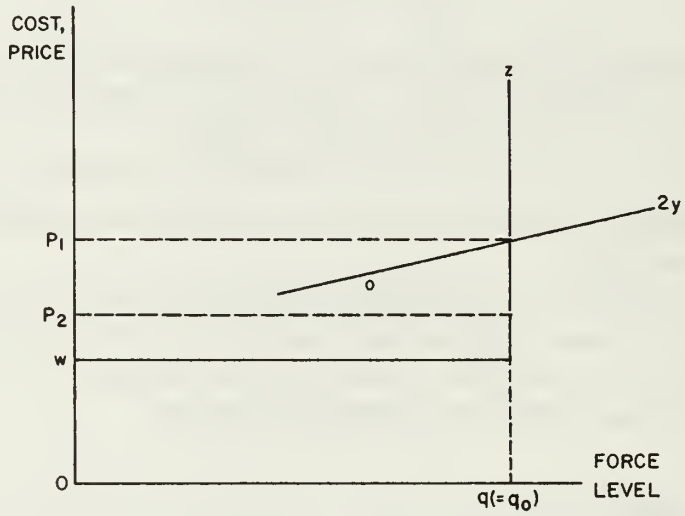

FIGURE 6. Both systems bearing common costs.

THE PROPOSED APPROACH TO PRICE DETERMINATION AND COST ALLOCATION

Military cost-effectiveness analysis generally takes one or another of the following forms: (a) maximize effectiveness subject to a cost constraint, or (b) minimize cost subject to achieving a given level of effectiveness. In this paper, attention will be devoted to the latter, but the basic approach is also applicable to the former.

As seen from the discussion above, it is critical to be able to set the prices of various missions. At first, this requirement appears unorthodox and difficult; but, in fact, it is not far from existing practice. In the case of a single system, the price of one unit of the system (say one plane), plus the wages of personnel, the expenditure for fuel, and whatever else is necessary for the unit to perform its mission, can be considered to be the price of the mission. Expressed somewhat diffe ently, the amount of expenditure necessary to meet a given mission requirement divided by the units of a system which must be used is the price of that mission as performed by a particular system.

The Alternative System Method

Consider two single-mission systems which perform different missions. The expenditures on these two systems are shown in Figures 7 and 8. In order to meet effectiveness requirements for each, q_1 and q_2 units are required respectively. To meet these requirements, expenditures of $Op_1c_1q_1$ and $Op_2c_2q_2$ are necessary, implying mission prices of p_1 and p_2.

If a third system performed both missions jointly, it might be procured in such quantity (q_3) that it met the requirements of mission 1 and part of the requirements of mission 2. This is shown in Figure 9. If we assume that the operating costs of system 3 are identical to those of systems 1 and 2, and the investment costs are exactly the same for system 3 as for the sum of systems 1 and 2, then, at the optimum, the prices of the missions for the multi-mission system are equal to those of the single-mission systems.

$$p_1 + p_2 = 2\ (x_3 + w).$$

This follows from (12), and allocation of common costs (investment costs in this case) follows (13). That is, in this particular case the shares of common costs differ solely because of the difference in mission prices.

FIGURE 7. Average price and required units of system in Mission 1 (alternative system method).

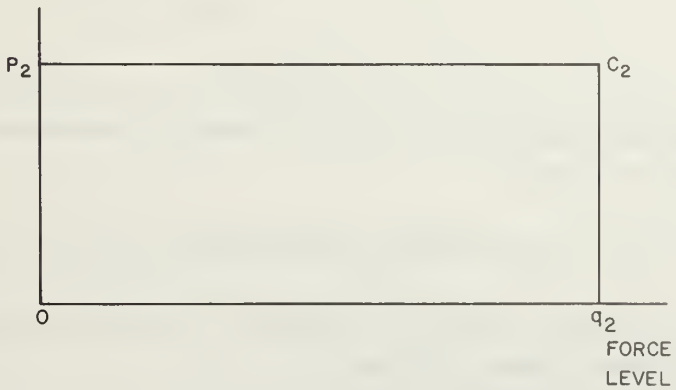

FIGURE 8. Price and required units in Mission 2 (alternative system method).

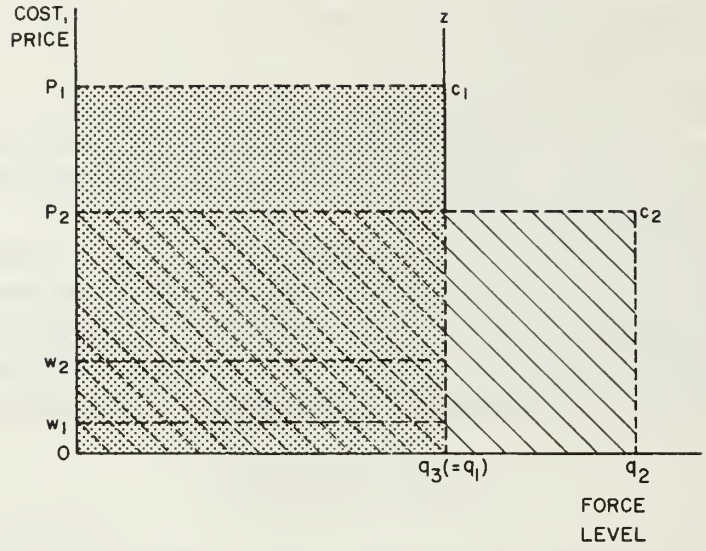

FIGURE 9. Costs and prices of both missions with a single system.

Clearly the assumptions of identical costs is unrealistic. There is no reason to assume that a multi-mission system will cost precisely the same as the sum of two single-mission systems. In fact, if it did there would be no apparent reason to buy it. The only apparent reason for a multi-mission system is that it meets a set of requirements less expensively than single-mission alternatives. Thus, to preserve the equality of prices and marginal costs for the multi-mission system, it is necessary to scale either the marginal costs or the prices. Since it makes no difference which is chosen, prices are scaled:

$$p_{13}+p_{23}=n(p_1+p_2)=2(x_3+w),$$

where p_{13} and p_{23} are the prices of missions 1 and 2 as performed system 3, and

$$n=\frac{2(x_3+w)}{(x_1+w)+(x_2+w)}, \tag{14}$$

from the conditions

$$p_1=x_1+w \text{ and } p_2=x_2+w,$$

established in (5). Thus, the proportions of common investment cost (for example) allocated to different missions are the respective terms of

$$(np_1-w)/2x_3+(np_2-w)/2x_3=1.$$

Next, consider the systems where certain investment and operating costs can be attributed to specific missions. The following notation will be used:

p_{13} = price of mission 1 performed by system 3

p_{23} = price of mission 2 performed by system 3

w_{13} = an approximation of marginal operating costs for mission 1 performed by system 3 (see definition of w prior to Equation (1)).

w_{23} = an approximation of marginal operating costs of mission 2 performed by system 3.

w_3 = an approximation of marginal operating costs of system 3 common to missions 1 and 2.

z_{13} = marginal operating cost of mission 1 performed by system 3

z_{23} = marginal operating cost of mission 2 performed by system 3

x_{13} = marginal investment cost of mission 1 performed by system 3

x_{23} = marginal investment cost of mission 2 performed by system 3

x_3 = marginal investment cost of system 3 common to missions 1 and 2

The critical condition for optimum resource allocation in the simple case presented in (11) was

$$z_1 + z_2 = 2x + w_1 + w_2.$$

If there are investment costs and operating costs common to each mission, as well as investment costs and operating costs specific to both missions, the specific costs are directly attributable to their respective missions. The condition can be written as

$$z_{13} + z_{23} = 2(x_3 + w_3) + x_{13} + w_{13} + x_{23} + w_{23},$$

which preserves the equality of short- and long-run costs for the system. Since the condition for optimum resource allocation is that output is such that price equals marginal cost,

$$p_{13} = z_{13} \text{ and } p_{23} = z_{23}.$$

Therefore,

(15) $$p_{13} + p_{23} = 2(x_3 + w_3) + x_{13} + w_{13} + x_{23} + w_{23}$$

and

$$(p_{13} - w_{13}) + (p_{23} - x_{23} - w_{23}) = 2(x_3 + w_3)$$

which implies that allocation of common investment and operating costs follows

(16) $$[(p_{13} - x_{13} - w_{13})/2(x_3 + w_3)] + [(p_{23} - x_{23} - w_{23})/2(x_3 + w_3)] = 1.$$

Of course, in the case where p_1 and p_2 are given from single-mission systems, then

$$p_{13} = np_1 \text{ and } p_{23} = np_2$$

from (14).

Knowledge of Historical or Simulated Trade-offs

If historical data, e.g., Viet Nam, Korea, or World War II, is relevant to the missions in question, perhaps some trade-offs can be established. Force structure analysis might also be useful in establishing such trade-offs. For example, suppose that as a result of such analysis, a trade-off could be established such that the outcome of a campaign would have been the same regardless of whether an amount of "output" m_1 of mission 1 or m_2 of mission 2 were provided. Therefore the price of mission 2 relative to mission 1 is

$$\frac{p_2}{p_1} = \frac{m_1}{m_2} = k$$

or

$$p_2 = kp_1.$$

The price of mission 1 performed by system 3 must be such that (15) will hold, that is

$$p_{13} + kp_{13} = 2(x_3 + w_3) + x_{13} + w_{13} + x_{23} + w_{23},$$

which implies

$$p_{13} = [2(x_3 + w_3) + x_{13} + w_{13} + x_{23} + w_{23}]/\ (1 + k). \tag{17}$$

Since $p_{23} = kp_{13}$, both p_{13} and p_{23} are determined and allocation of common costs follows (16).

Expert Opinion

Expert opinion may be used to establish trade-offs needed for common cost allocation. This appears to be the implicit assumption underlying the two-ship method, in which judgment must be made as to which mission is "major" and which is "minor." The principal distinction is that under the proposal of this paper, if two missions are believed to be approximately equal in importance, common costs would be allocated according to relative prices, taking incremental investment costs and operating costs into account, rather than allocating all or none of the common costs to each system. If one was thought to be slightly more important for the system in question than the other, relative prices might be set such that $p_1/p_2 = 55/45$ and so forth. The setting of an absolute price for one mission would follow (17) and the allocation of common costs would follow (16).

A Simple Example of Mission and System Comparisons

At the beginning of the paper, Table 1 presented the costs of achieving given amounts of output in three missions by five different systems—three single mission and two multi-mission. To illustrate the method, consider the basic problems of the paper: (1) which systems are least expensive in carrying out their multiple missions, and (2) what are the costs of multiple missions supplied by a single system.

Since we have provided a single-mission system for each of our alternatives, we use the alternative system method of determining mission price. The information of Table 1 and the results of allocation are presented in Table 2. In the illustration, it will be assumed that the entries in the table are costs per unit and that marginal investment and operating costs are constant. Thus, the entries are approximations of marginal costs, i.e., the w's and x's in the notation above. The functions, such as (6a) and (6b) which would yield these parameters could be developed from statistical cost estimating, industrial engineering studies, analogy or even expert opinion, and should reflect all of the sources of costs, e.g., equipment, fuel, personnel, etc. that make the system operational.

Beginning with System D, allocation follows from (14) and (16). That is,

$$n_D = \frac{2x_D + w_{D1} + w_{D2}}{(x+w)_A + (x+w)_B} \quad \text{or} \quad \frac{100+20+30}{65+45} = 1.36,$$

TABLE 2. *Marginal Costs of Missions 1, 2, and 3 for all Systems Before and After Allocation of Common Costs*

Costs	Before allocation					After allocation	
	Systems					Systems	
	A	B	C	D	E	D	E
Common				50	60		
Mission 1	65			20		55	
Mission 2		45		30	30	45	49
Mission 3			60		10		51

and

$$n_D(p_A + p_B) = 2x_D + w_{D1} + w_{D2} \quad \text{or} \quad 1.36\,(65+45) = 100+20+30,$$

which implies

$$(n_D p_A - w_{D1})/2x_D + (n_D p_B - w_{D2})/2x_D = 1, \quad \text{or} \quad 0.69 + 0.31 = 1.00.$$

That is, 69 percent of the common cost is borne by Mission 1 and 31 percent by Mission 2. The long run marginal costs, with the common costs allocated to the specific missions, are therefore:

$$Y_{D1} = 0.69(50) + 20 = 55 \text{ and } Y_{D2} = 0.31(50) + 30 = 45.$$

The same procedure is followed for System E, where

$$n_E = \frac{2x_E + w_{E2} + w_{E3}}{(x+w)_B + (x+w)_C} = \frac{2(60)+30+10}{45+60} = 1.52,$$

and

$$(n_E p_B - w_{E2})/2x_E + (n_E p_C - w_{E3})/2x_E = 1 \qquad \text{or} \qquad 0.32 + 0.68 = 1.$$

The shares of common cost allocated to Missions 2 and 3, 32 and 68 percent, are then applied to the marginal common costs, and the sums of the allocated marginal common cost plus the incremental cost for each system are

$$Y_{E2} = 0.32(60) + 30 = 49 \text{ and } Y_{E3} = 0.68(60) + 10 = 51.$$

Now that the costs have been allocated, let us consider the results. Taking System D first, we see that its cost for performing Mission 1 is less than for the single-mission system, System A. Its cost of performing Mission 2 is exactly the same as that of System B. Therefore, it would appear that System D is worthy of consideration for procurement and deployment for Missions 1 and 2, since its cost in each mission is, at worst, no greater than that of competing systems.

Turning to System E, we find that it is more expensive in Mission 2 than are Systems B and D, i.e., its incremental cost exceeds its price. On the other hand, it is less expensive than C in its performance of Mission 3. Since its long-run marginal cost for Mission 2 exceeds that of B and D, System 3 will never be used in that mission. This implies that *all* common costs are to be allocated to missions which it does perform, Mission 3 in this case. However, this means that the long-run marginal cost of Mission 3 is now 70 instead of 51, and it too now exceeds the cost of the single-mission alternative. Thus, it will never be used according to our simplified analysis and does not appear to be a good candidate.

This result has a paradoxical element in that even though E's marginal costs exceed those of the single mission alternatives when considered separately, the sum of the marginal cost is *lower* than the sum of the single-mission alternatives. This, of course, is due to all costs being allocated to Mission 3 since Mission 2 is not performed. Does the allocation of common costs thus lead us astray? Would we not be more likely to make good decisions on systems use and procurement if we simply compared total long-run marginal costs and ignore allocation? The answer in this particular example, at any rate (and probably generally, too) is no. Consider the possible ways of accomplishing Missions 1, 2, and 3. The combinations and their associated costs are presented in Table 3. The results of considering al-

TABLE 3. *Marginal Costs of Alternative Combinations of Systems Providing Three Missions*

Combination	Marginal costs of systems					
	A	B	C	D	E	Total
A, B, C	65	45	60			170
A, E	65				100	165
C, D			60	100		160
D, E				100	100	200

ternative combinations is that the least-cost combination is C and D. In the case of a direct comparison with Systems B and C alone, however, System E does have an advantage. In this case, what is implied is that cost penalities be accepted in Mission 2 in order to retain low-cost capabilities in Mission 3. In this

case, the marginal cost of Mission 2 would be set exactly equal to the price of the alternative system, reducing it from 49 to 45, and the marginal cost of Mission 3 would be raised from 51 to 55, i.e., four units of marginal common cost would be reallocated from Mission 2 to Mission 3.

Common Cost Allocation, Time and Discount Rates

Investments in military systems, like other public and private systems, have particular useful lives and are subject to particular rates of discount. Exactly how useful lives and discount rates are determined are difficult problems in their own right and will not be discussed here. Suffice it to say that useful lives are functions of wear and tear and obsolescence, and discount rates reflect the terms under which the values of present and future costs and effectiveness are compared.

Several measures have been employed for evaluating benefits and costs over time. Although each has its strengths and weaknesses, the best general measure appears to be the net discounted present value of an investment.* This is the measure to be employed for the purpose of illustrating how discounting and system life may be handled for multiple-mission systems. First, let us consider a single-mission system. Its net discounted present value (V) is:

$$V = \Sigma_i d_i (pq_0 - C_o), \tag{12}$$

where d_i is the discount rate in year i, or $\frac{1}{(1+r)^i}$ where r is the rate of interest in use for military systems and is assumed, for the sake of simplicity, to be constant for all relevant years.† The other variables are as described above. In words, then, V is the sum of a stream of net benefits, each year's entry being discounted by a greater amount than that of the year before.**

If we extend this to the case of two multiple-mission systems where the investment costs are common to both systems and have identical useful lives and are subject to the same interest rates, we have

$$V_{12} = \Sigma_i d_i \ (p_1 q_0 + p_2 q_0 - C_f(q_0) + w_1 q_o + w_2 q_o). \tag{18}$$

The first order condition for the maximization of the net discounted present value of the two missions performed by the system is

$$\frac{\partial V_{12}}{\partial q_o} = 0.$$

This implies, following Boiteux [2, appendix],

$$\Sigma_i d_i \ (p_1 + p_2) = \Sigma_i d_i \ (2x + w_1 + w_2). \tag{19}$$

Since both sides may be divided by $\Sigma_i d_i$, we see that (19) reduces to

$$p_1 + p_2 = 2x + w_1 + w_2.$$

*For a comparison of the more prominent measures, see Baumol [1. ch. 19].

†The interest rate is presumably based on some notion of social time preference or opportunity cost. For a discussion of some of the issues involved see (Prest and Turvey, [7, pp. 697–700].

**The link from "benefits" to expenditure (pq_o) in our context is that pq_o is the expenditure necessary to meet a particular requirement, and it is the meeting of the requirement that is the benefit.

Thus, allocation in this case follows the same lines as (13), (16), etc.; and the interest rate and useful life plays no role.

If these are different useful lives of the system in different missions, and/or the interest rate differs, a general solution of the allocation problem has not been found, although solutions have been found for specific cost functions, e.g., linear functions. The problem arises from the translation of units of output of particular missions to units of the multimission systems. The simplicity of the cost-allocation scheme proposed in this paper is due to Boiteux's demonstration that this can be done with great generality when time does not enter the picture. This breaks down, however, where the missions have different useful lives while performed by the multiple-mission system in question or have different interest rates, since it is not generally possible to translate them from specific missions to units of the multiple-mission system. It appears that the basic approach retains its validity, but not its simplicity.

CONCLUSION

This paper has presented a new technique for allocating the common costs of multiple-mission systems. One major departure from existing practice is that the basis for allocation is to be found in the importance of the missions as reflected in their relative prices or, more generally, on an assessment of the relative abilities of a system to carry out alternative missions. The second major departure is that it uses marginal conditions rather than proportional or "either-or" allocations. Thus, unlike existing techniques, it is consistent with the principles of efficient resource allocation.

These are not as drastic departures from current practice as they may seem, since some notion of relative importance is implicit in the distinction of major from minor missions in the currently employed two-ship method of allocation. The two-ship method also adheres to a rough approximation to marginal principles. In a very real sense, what has been presented above can be regarded as a generalization of the two-ship method, as well as an explicit statement of the principles underlying it.

The major problem in employing the proposed method is, of course, to develop means of measuring the relative prices of different missions supplied by a system. Several tentative suggestions have been offered which, crude as they are, should aid in achieving better allocation of common costs. In all likelihood better means can and will be devised through experience if the proposed approach is employed.

ACKNOWLEDGMENT

The author is grateful for discussions and correspondence with A. S. Rhode, J. T. Kammerer, K. F. Linder, Saul Gass, George Taylor, Kenneth Babka and William Baumol. I wish to thank them but also absolve them of any blame for whatever errors or misconceptions remain.

REFERENCES

[1] Baumol, W. J., *Economic Theory and Operations Analysis* (Englewood Cliffs, N.J., Prentice-Hall, 1965), (2nd ed.).

[2] Boiteux, M., "Peak Load Pricing" (J. R. Nelson, ed), *Marginal Cost Pricing in Practice* (Prentice-Hall, Englewood Cliffs, N.J., 1964), chap. 4, pp. 59–90.

[3] Carlson, S., *A Study on the Pure Theory of Production* (Kelley and Millman, New York, 1956).

[4] Crow, R. T., "The Allocation of Common Costs of Multiple-Mission Systems," a report to Systems Analysis, Chief of Naval Operations, Contract No. N00014–70–C–0086 by MATHEMATICA, Inc., Bethesda, Md. (Nov. 1971).

[5] Grey, J. C., *Cost Analysis Methodology* (Fire Support Study Working Paper No. 9), Dahlgran, Va.: U.S. Naval Weapons Laboratory (July 1970).

[6] Littlechild, S. C., "Marginal Cost Pricing with Joint Costs," Economic Journal *LXXX*, 323–335 (June 1970).

[7] Prest, A. R. and R. Turvey, "Cost-Benefit Analysis: A Survey," Economic Journal *LXXV*, 683–735 (Dec. 1965).

AN EXPLICIT GENERAL SOLUTION IN LINEAR FRACTIONAL PROGRAMMING*

A. Charnes

Center for Cybernetic Studies
University of Texas

W. W. Cooper

School of Urban and Public Affairs
Carnegie-Mellon University

ABSTRACT

A complete analysis and explicit solution is presented for the problem of linear fractional programming with interval programming constraints whose matrix is of full row rank. The analysis proceeds by simple transformation to canonical form, exploitation of the Farkas-Minkowski lemma and the duality relationships which emerge from the Charnes-Cooper linear programming equivalent for general linear fractional programming. The formulations as well as the proofs and the transformations provided by our general linear fractional programming theory are here employed to provide a substantial simplification for this class of cases. The augmentation developing the explicit solution is presented, for clarity, in an algorithmic format.

I. INTRODUCTION

The linear fractional programming problem arises in many contexts with relatively simple constraint sets, e.g., in the reduction of integer programs to knapsack problems, in attrition games, and in Markovian replacement problems as well as in Neyman-Pearson rejection region selection problems. Illustrative examples are provided by G. Bradley [5] or F. Glover and R. E. Woolsey [12]†, J. Isbell and W. Marlow [13], C. Derman [10], and M. Klein [16].

The linear fractional programming problem in all generality, and with all singular cases considered, was reduced in [8] to at most a pair of ordinary linear programming problems. This immediately made available all of the algorithms, interpretations, etc., that are associated with linear programming. This includes, we should note, access to any ordered field,** and any of the algorithms and the computer codes for linear programming problems which, by virtue of [8], thereby also become available for any problem in linear fractional form. Thus, with the development in [8], the work in linear fractional programming took a different form from its previous sole concern with the development of special types of algorithms for dealing with this kind of problem.

---

*This research was partly supported by a grant from the Farah Foundation and by ONR Contracts N00014-67-A-0126-0008 and N00014-67-A-0126-0009 with the Center for Cybernetic Studies, The University of Texas. This report was also prepared as part of the activities of the Management Sciences Research Group at Carnegie-Mellon University under Contract N00014-67-A-0314-0007 NR 047-048 with the U.S. Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

† See also E. Balas and M. Padberg [1].

** See, e.g., the development of the opposite sign theorem and related developments in [7].

In the present paper, we apply our reduction, as given in [8], to a general class of linear fractional problems, viz., those for which the constraint set is given by

$$a \leqq Ax \leqq b \tag{1.1}$$

so that this part of the model is in "interval programming" form.* Here we shall assume that the matrix A is full row rank and the vectors a, b, and x meet the usual conditions for conformance. This means that the constraint set is a parallelopiped. See the Final Appendix in [7].

Subject to conditions (1.1), we wish to

$$\text{maximize } R(x) \equiv \frac{N(x)}{D(x)} \equiv \frac{c^T x + c_0}{d^T x + d_0} \not\equiv \text{constant}, \tag{1.2}$$

so that we are now concerned with a problem of linear fractional programming. Because A is of full row rank it has a right inverse, $A^\#$, and hence we can write

$$AA^\# = I. \tag{1.3}$$

Now, setting

$$y = Ax$$

or

$$x = A^\# y + Pz,$$

where

$$P \equiv I - A^\# A \tag{1.4}$$

and

$$z \text{ is arbitrary,}$$

we obtain

$$\max R(y, z) = \frac{c^T A^\# y + c^T P z + c_0}{d^T A^\# y + d^T P z + d_0},$$

subject to

$$a \leqq y \leqq b \tag{1.5}$$

in place of (1.1) and (1.2).

Because z is arbitrary,** unless

$$c^T P = d^T P = 0 \tag{1.6}$$

---

*See [2]–[4] and [18]–[19].

**Observe that we have ruled out the case in which R is identically constant in (1.2).

we shall obtain max $R = \infty$. In order to avoid repetitious arguments, however, we defer the proof of this until we have discussed the situation $c^TP = d^TP = 0$. See section IV, below.

Waiving this consideration, we shall next proceed to solve this problem explicitly and in all generality by means of the following three characterizations: The denominator $D(x)$ is (1) bisignant, (2) unisignant and nonvanishing, or (3) unisignant and vanishing on the constraint set. In (1), i.e., the bisignant case, we shall show that $R(x) = \infty$. Furthermore, we shall show how to identify this case at the outset so that it may be discarded from further consideration. This will leave us with only cases (2) and (3) to examine where we shall proceed to transformations from which a one-pass numerical comparison of coefficients makes explicit the optimal value and solution.

After this has all been done, we shall then return to assumption (1.6) in a way that utilizes the preceding developments. Finally we shall supply numerical examples to illustrate some of these situations and then we shall draw some conclusions for further research which will return to the remarks at the opening of this section.

II. BISIGNANT DENOMINATORS

Employing assumption (1.6), our problem is

$$\max R(y) = \frac{c^TA^*y + c_0}{d^TA^*y + d_0}, \tag{2.1}$$

subject to

$$a \leqslant y \leqslant b$$

in place of (1.5). Note, however, that here, and in the following, we shall slightly abuse notation by continuing to use the symbols R, N, D, as in (1.1) and (1.2) even though we mean the transformed function, as in (2.1).

Let $\overset{\vee}{D}, \hat{D}$ denote the maximum and minimum, respectively, of D over the constraint set. We note:

LEMMA 1:

(a) D is bisignant if and only if $\overset{\vee}{D} > 0$ and $\hat{D} < 0$.

(b) D is unisignant if and only if either $\overset{\vee}{D} \leqslant 0$ or $\hat{D} \geqslant 0$.

In terms of y, since we can choose each component of y independently (see (1.5)), we can express $\overset{\vee}{D}$ and $\hat{D}$ immediately as

$$\begin{aligned} \overset{\vee}{D} &= \sum_{+} d'_j b_j + \sum_{-} d'_j a_j + d_0 \equiv \max D(y) \\ \hat{D} &= \sum_{+} d'_j a_j + \sum_{-} d'_j b_j + d_0 \equiv \min D(y), \end{aligned} \tag{2.2}$$

where $d'_j \equiv (d^TA)_j$ is the jth element of d^TA^* and "+" or "—" indicates that the summation is on only the positive or negative d'_j.

Let us first consider the bisignant situation. If we make the transformation of variables

$$\begin{aligned} y_j - a_j &= k_j\zeta_j, \; d_j \geqslant 0 \\ b_j - y_j &= k_j\zeta_j, \; d_j < 0, \end{aligned} \tag{2.3}$$

where the $k_j > 0$ will be suitably chosen, the constraints transform to

$$0 \leq k_j \zeta_j \leq b_j - a_j$$

(2.4) or

$$0 \leq \zeta_j \leq \delta_j \equiv \frac{b_j - a_j}{k_j}.$$

Without loss of generality, the δ_j are positive, since otherwise $\zeta_j = 0$ and does not enter into the optimization.

By choosing the k_j suitably, we obtain the form

$$R(\zeta) = \frac{\sum_j \gamma_j \zeta_j + \gamma_0}{\sum_+ \zeta_j + \sum_- \zeta_j - 1}. \tag{2.5}$$

Note that $\sum_+ \delta_j + \sum_- \delta_j > 1$, since otherwise D would not be bisignant.

One of the following two cases must now hold:

CASE (i): for some $\bar{\zeta}$, $0 \leq \bar{\zeta} \leq \delta$, such that

$$D(\bar{\zeta}) = 0 \text{ we have } N(\bar{\zeta}) \neq 0,$$

or else

CASE (ii): for every ζ, $0 \leq \zeta \leq \delta$, such that

$$D(\zeta) = 0 \text{ we have } N(\zeta) = 0.$$

In Case (i), since $N(\zeta)$ is continuous, there is a neighborhood of $\bar{\zeta}$ in which $N(\zeta)$ is unisignant. Since

$$\sum_+ \bar{\zeta}_j + \sum_- \bar{\zeta}_j = 1 < \sum_+ \delta_j + \sum_- \delta_j, \tag{2.6}$$

and $0 \leq \zeta_j \leq \delta_j$ in the constraint set, we can choose $\epsilon_j \geq 0$, $\sum_j \epsilon_j > 0$, so that $0 \leq \bar{\zeta} \pm \epsilon \leq \delta$, $\operatorname{sgn} N(\bar{\zeta} + \epsilon) = \operatorname{sgn} N(\bar{\zeta} - \epsilon)$ and $D(\zeta + \epsilon) > 0 > D(\bar{\zeta} - \epsilon)$. By approaching $\bar{\zeta}$ along the line segment from one of $\bar{\zeta} + \epsilon, \bar{\zeta} - \epsilon$, we can make $R(\zeta) \to \infty$.

In Case (ii), we must have $\gamma_j = 0$ for all j such that $d_j = 0$. For $D(\zeta) = 0$ involves specifying only the ζ_j for $d_j > 0$ and $d_j < 0$. If $\gamma_{j_0} \neq 0$ for $d_{j_0} = 0$, then having made $D(\zeta) = 0$, we can change the value of $N(\zeta)$ by changing ζ_{j_0}. Thus $D(\zeta) = 0$ would *not* imply $N(\zeta) = 0$. We therefore drop the "+", "−" notation is considering Case (ii) and rewrite it as $\sum_j \gamma_j \zeta_j + \gamma_0 = 0$ whenever $0 \leq \zeta_j \leq \delta_j$ and $\sum_j \zeta_j = 1$.

By letting $y_j = \zeta_j y_0$, $y_0 \geq 0$, this becomes $\pm\left(\sum_j \gamma_j y_j + \gamma_0 y_0\right) \geq 0$, whenever

$$\sum_j y_j - \quad y_0 = 0$$

(2.7)
$$\begin{aligned} -y_j + \delta_j y_0 &\geq 0 \\ y_j \qquad\quad &\geq 0 \\ y_0 &\geq 0. \end{aligned}$$

Note that the implication extends to $y_0 = 0$ since $y_0 = 0$ implies *all* $y_j = 0$.

We now apply the Farkas-Minkowski Lemma* to the pair of implications in (2.7) and obtain

(2.8)
$$\begin{cases} \gamma_j = \mu^+ - \theta_j^+ + \nu_j^+ \\ \gamma_0 = -\mu^+ + \sum_j \delta_j \theta_j^+ + \nu_0^+; \qquad \theta_j^+, \nu_j^+, \nu_0^+ \geq 0, \end{cases}$$

for the first implication, viz., $\left(\sum_j \gamma_j y_j + \gamma_0 y_0\right) \geq 0$. For the second one viz., $\left(\sum_j \gamma_j y_j + \gamma_0 y_0\right) \geq 0$, we obtain,

(2.9)
$$\begin{cases} -\gamma_j = \mu^- - \theta_j^- + \nu_j^- \\ -\gamma_0 = -\mu^- + \sum_j \delta_j \theta_j^- + \nu_0^-; \qquad \theta_j^-, \nu_j^-, \nu_0^- \geq 0. \end{cases}$$

Adding the first expressions in (2.8) and (2.9),

$$0 = \mu^+ + \mu^- - (\theta_j^+ + \theta_j^-) + (\nu_j^+ + \nu_j^-)$$

(2.10) or

$$\theta_j^+ + \theta_j^- = (\mu^+ + \mu^-) + (\nu_j^+ + \nu_j^-).$$

Adding the second pair

$$0 = -(\mu^+ + \mu^-) + \sum_j \delta_j (\theta_j^+ + \theta_j^-)(\nu_0^+ + \nu_0^-)$$

(2.11) or

$$\mu^+ + \mu^- = \sum_j \delta_j (\theta_j^+ + \theta_j^-) + \nu_0^+ + \nu_0^-.$$

Since each term on the right is nonnegative, we have

(2.12)
$$\mu^+ + \mu^- \geq 0.$$

Next, substituting from (2.10) into (2.11), we get

$$0 = -(\mu^+ + \mu^-) + \sum_j \delta_j (\mu^+ + \mu^-) + \sum_j \delta_j (\nu_j^+ + \nu_j^-) + \nu_0^+ + \nu_0^-$$

(2.13) or

$$0 = \left(\sum_j \delta_j - 1\right)(\mu^+ + \mu^-) + \sum_j \delta_j \nu_j^+ + \sum_j \delta_j \nu_j^- + \nu_0^+ + \nu_0^-.$$

Since the right-hand side is a sum of nonnegative terms, each of these must be zero. Moreover,

$$\mu^+ + \mu^- = 0 \qquad \text{since } \sum_j \delta_j - 1 > 0$$

---

*See Appendix C in [7].

(2.14) $$\nu_j^+ = \nu_j^- = 0 \qquad \text{since } \delta_j > 0$$

$$\nu_0^+ = \nu_0^- = 0.$$

By virtue of (2.14), and going back to (2.10),

(2.15) $$\theta_j^+ + \theta_j^- = 0.$$

Further, with $\theta_j^+, \theta_j^- \geq 0$, we must have $\theta_j^+ = \theta_j^- = 0$ for all j. Therefore, $\gamma_j = \mu^+$, for all j, and $\gamma_0 = -\mu^+$, so that we have

(2.16) $$R(\zeta) = \frac{\sum_j \gamma_j \zeta_j + \gamma_0}{\sum_j \zeta_j - 1} = \frac{\mu^+\left(\sum_j \zeta_j - 1\right)}{\sum_j \zeta_j - 1}$$

$$= \mu^+ = \text{constant}.$$

In other words, Case (ii) can only occur in the trivial instance where the numerator is a constant multiple of the denominator. In this case, each coefficient in the numerator is the same multiple of the corresponding coefficient in the denominator, and this would have to be true in the original $N(x)$, $D(x)$ description and hence obvious upon comparing the initial coefficients. Since we have ruled out this very obvious case (see (1.2)), we have only $\max R(x) = \infty$ when $D(x)$ is bisignant on the constraint set.

III. UNISIGNANT DENOMINATORS

The unisignant cases now remain to be considered. If $\overset{\vee}{D} \leq 0$ we multiply both N and D by -1, (thus not altering the value of R) and we are then reduced to "D" ≥ 0. With this normalization, we make a transformation of variables as in (2.3),

(3.1) $$\begin{aligned} y_j - a_j &= g_j \xi_j, \qquad d_j \geq 0 \\ b_j - y_j &= g_j \xi_j, \qquad d_j < 0, \end{aligned}$$

where the $g_j > 0$ will be suitably chosen. The constraint set will now be

(3.2) $$0 \leq \xi_j \leq \delta_j = \frac{b_j - a_j}{g_j},$$

and, first considering the case where $\hat{D} > 0$, the g_j can be chosen so that the problem is

(3.3) $$\max R(\xi) = \frac{\sum_j \gamma_j \xi_j + \gamma_0}{\sum_j \xi_j + 1}, \qquad 0 \leq \xi_j \leq \delta_j.$$

In (3.3) the summation is only over "+" and "−" because, *the denominator being positive*, optimal values for the ξ_j such that $d_j=0$ can be specified as $\xi_j=0$ when $\gamma_j<0$; $\xi_j=\delta_j$ when $\gamma_j\geq 0$, and these new constant terms are assumed to be already contained in γ_0. By the reduction that we gave in [8], however, the equivalent linear programming problem is

$$\max \sum_j \gamma_j\eta_j+\gamma_0\eta_0,$$

subject to

$$\sum_j \eta_j+\eta_0=1$$

(3.4)

$$\eta_j-\delta_j\eta_0\leq 0$$

$$\eta_j\geq 0,$$

where we can also note that these constraints imply $\eta_0>0$.

The dual to (3.4) is

$$\min u$$

subject to

$$u+\omega_j\geq\gamma_j$$

(3.5)

$$u-\sum_j \delta_j\omega_j=\gamma_0$$

$$\omega_j\geq 0.$$

We shall employ this dual in an essential manner to obtain our desired one-pass argument for obtaining an optimum. At each step in the procedure, we shall have a solution to a less restrictive problem than the dual problem and an associated primal feasible solution.

Suppose the γ's are renumbered so that $\gamma_1\geq\gamma_2\geq\ \ldots\ \gamma_n$;

Then Case (i) $\gamma_0\geq\gamma_1$ has the immediately obvious primal solution $\eta_0^*=1$, $\eta_j^*=0$, and $\max R(\xi)=\gamma_0$.

In the contrary case, Case ii,

$$\gamma_1\geq\ \ldots\ \geq\gamma_p>\gamma_0\geq\gamma_{p+1}\geq\ \ldots\ \geq\gamma_n,$$

we build up an algorithm based on the dual problem in which we choose u^q at the qth step to satisfy

$$u^q+\omega_j^q=\gamma_j,\qquad j=1,\ \ldots,q$$

(3.6)

$$u^q-\sum_{j=1}^{q}\delta_j\omega_j^q=\gamma_0.$$

Using the first q equations to obtain ω_j^q in terms of u^q and substituting in the last equation, we obtain

$$u^q\left(1+\sum_1^q \delta_j\right)=\sum_1^q \gamma_j\delta_j+\gamma_0 \tag{3.7}$$

and hence

$$u^q=\frac{\sum_1^q \gamma_j\delta_j+\gamma_0}{1+\sum_1^q \delta_j}. \tag{3.8}$$

Thus, u^q is a convex combination of $\gamma_0, \gamma_1, \ldots, \gamma_q$ with proportionality constants $1, \delta_1, \ldots, \delta_q$. Note that if $u^q \leqq \gamma_q$, then $u^q, \omega_j^q, j=1, \ldots, q$, satisfy the first q constraints plus the "γ_0" constraint of the dual problem; hence satisfy a less restrictive problem than the dual.

If we take

$$\eta_0^q=1 \Big/ \left(1+\sum_1^q \delta_j\right) \tag{3.9}$$

and

$$\eta_j^q=\delta_j \Big/ \left(1+\sum_1^q \delta_j\right), \qquad j=1, \ldots, q \tag{3.10}$$

$$\eta_j^q=0, \qquad j>q,$$

then $\eta_j^q, j=0, \ldots, n$ is a feasible solution to the primal problem and $\gamma^T\eta^q+\gamma_0\eta_0^q=u^q$ (by substitution in (3.4) and comparison with (3.8)).

Hence, *whenever we can get* u^q, ω_j^q *feasible for the dual problem, we will have a primal feasible solution for* η^q *with the same functional value and thus we will have an optimal pair of dual solutions.* This, plus the equivalences maintained via (3.1) and our theory from [8], thus justifies the development that we detail as follows:

To start,

$$u^1=(\gamma_0+\gamma_1\delta_1)/(1+\delta_1). \tag{3.11}$$

(Note, $u^1>\gamma_0$ since $\gamma_1>\gamma_0$ and u^1 is a proper convex combination of γ_1, γ_0.)

We check:

Is $$u^1 \geqslant \gamma_2?$$

If yes:

we are done:

$$u^1=u^*$$

$$\omega_j^1=\omega_j^*=\begin{cases}\gamma_j-u^*, & j=1\\ 0, & j>1\end{cases}$$

and

$$\eta_0^1 = 1/1 + \delta_1$$

$$\eta_1^1 = \eta_1^* = \delta_{1/1+\delta_1}$$

$$\eta_j^1 = \eta_j^* = 0,\ j > 1.$$

If No:

$$u^1 < \gamma_2,$$

then

$$u^2 = \frac{\gamma_0 + \gamma_1\delta_1 + \gamma_2\delta_2}{1 + \delta_1 + \delta_2}$$

$$= \left(\frac{\gamma_0 + \gamma_1\delta_1}{1 + \delta_1}\right)\left(\frac{1 + \delta_1}{1 + \delta_1 + \delta_2}\right) + \left(\frac{\delta_2}{1 + \delta_1 + \delta_2}\right)\gamma_2$$

$$= u^1\left(\frac{1 + \delta_1}{1 + \delta_1 + \delta_2}\right) + \left(\frac{\delta_2}{1 + \delta_1 + \delta_2}\right)\gamma_2 < \gamma_2,$$

since $u^1 < \gamma_2$.

Next,

$$\text{Is } u^2 \geqslant \gamma_3?$$

If yes: We are done with the substitutions indicated by (3.6).
If No:

$$u^2 < \gamma_3 \text{ and we continue to } u^3.$$

This process must stop by u^p at the latest since

$$u^p = \frac{\gamma_0 + \sum_1^p \gamma_j\delta_j}{1 + \sum_1^p \delta_j} > \gamma_0 \geqslant \gamma_{p+1} \geqslant \ldots \geqslant \gamma_n.$$

Thus $\max R(\xi) = u^s$ where s is the least positive integer such that $u^s \geqslant \gamma_{s+1}$, and

$$u^s = \frac{\gamma_0 + \sum_1^s \gamma_j\delta_j}{1 + \sum_1^s \delta_j}.$$

This concludes the case $\hat{D} > 0$.

The remaining case has $\hat{D}=0$. By making a transformation of variables as in (3.1), and choosing the $\xi_j=0$ suitably, we obtain

$$R(\xi) \equiv \frac{\sum_1^n \gamma_j \xi_j + \gamma_0}{\sum_1^n \xi_j}, \qquad 0 \leq \xi_j \leq \delta_j. \tag{3.12}$$

We may dispose of two situations immediately

(i) $\gamma_0 > 0$: then $R(\xi) \to \infty$ as $\xi_j \to 0$.

(ii) $\gamma_0 = 0$: here the dual problems are

$$\max \sum_1^n \gamma_j \eta_j \qquad\qquad \min u$$

$$\text{with } \sum_1^n \eta_j = 1 \qquad\qquad \text{with } u + \omega_j \geq \gamma_j$$

$$\eta_j - \delta_j \eta_0 \leq 0 \qquad\qquad -\sum_1^n \delta_j \omega_j \geq 0$$

$$\eta_j, \eta_0 \geq 0 \qquad\qquad \omega_j \geq 0.$$

An optimal solution pair is $\omega_j^* = 0$, $u = \gamma_1 \equiv \max \gamma_j$, and $\eta_1^* = 1$, $\eta_j^* = 0$, $j \geq 2$. The maximum of $R(\xi)$ is thus γ_1.

The remaining instance is

(iii) $\gamma_0 < 0 : \gamma_1 \geq \cdots \geq \gamma_p \geq \gamma_0 \geq \gamma_{p+1} \geq \cdots \geq \gamma_n$.

The dual linear programs can now be written

$$\max \sum_j \gamma_j \eta_j + \gamma_0 \eta_0 \qquad\qquad \min u$$

$$\text{with } \sum_j \eta_j = 1 \qquad\qquad \text{with } u + \omega_j \geq \gamma_j$$

$$\eta_j - \delta_j \eta_0 \geq 0 \qquad\qquad \sum_j \delta_j \omega_j = |\gamma_0|$$

$$\eta_j \geq 0 \qquad\qquad \omega_j \geq 0.$$

As before, we define

$$u^q + \omega_j^q = \gamma_j, \qquad j = 1, \ldots, q$$

$$\sum_1^q \delta_j \omega_j^q = |\gamma_0|.$$

This yields

$$u^q = \left(\sum_1^q \gamma_j \delta_j + \gamma_0 \right) \Big/ \sum_1^q \delta_j.$$

It may be easily verified that

$$u^q = u^{(q-1)} \left(\frac{\sum_1^{q-1} \delta_j}{\sum_j^q \delta_j} \right) + \left(\frac{\delta_j}{\sum_1^q \delta_j} \right) \gamma_q,$$

or, what is the same thing, u^q is a proper convex combination of $u^{(q-1)}$ and u^q. Thus, if $u^r < \gamma_{r+1}$, then $u^r < u^{r+1} < \gamma_{r+1}$ as in our earlier argument for $\hat{D} > 0$.

The steps of our process are as before: if $u^q \geq \gamma_{q+1}$ we are done; otherwise, we test u^{q+1} against γ_{q+2}. At worst we are done with

$$u^n = \left(\sum_1^n \gamma_j \delta_j + \gamma_0 \right) \Big/ \sum_1^n \delta_j.$$

As before, $u^s = \max R(\xi)$, where s is the least positive integer such that $u^s \geq \gamma_{s+1}$.

IV. $R(y, z)$

Returning to (1.5) we consider the remaining cases in which either

(a) $$d^T P = 0, \qquad c^T P \neq 0$$

(4) or

(b) $$d^T P \neq 0.$$

In case (a), since z is arbitrary we can make $\pm c^T P z \to \infty$, hence $\max R(y, z) \to \infty$. In case (b), we are in the bisignant denominator situation since we can make $d^T P z \to \pm \infty$. The argument of the bisignant section of this paper (with the additional variables, z) now shows that $\max R = \infty$ since we have ruled out, a priori, the case in which $R \equiv$ constant.

V. EXAMPLES

Some examples may help to fix and sharpen some of the preceding developments. Thus consider

$$\max R(x) \equiv \frac{3x_1 - x_3 + 4}{2x_2}$$

with

(5.1)
$$-1 \leq x_1 + x_3 \leq 2$$
$$1 \leq x_2 \qquad \leq 5,$$

and the variables x_1, x_2, x_3 are otherwise unrestricted.

Here we have,*

$$A=\begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 0\end{bmatrix}, \qquad A^{\#}=\begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}$$

$$P\equiv (I-A^{\#}A)=\begin{bmatrix}0 & 0 & -1\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \tag{5.2}$$

$$a=\begin{pmatrix}-1\\ 1\end{pmatrix}, \qquad b=\begin{pmatrix}2\\ 5\end{pmatrix}.$$

To exhibit the development in full detail we next write

$$c^TA^{\#}y=(3,0,-1)\begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix}=(3,0)\begin{pmatrix}y_1\\ y_2\end{pmatrix}=3y_1$$

$$c^TPz=(3,0,-1)\begin{bmatrix}0 & 0 & -1\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{pmatrix}z_1\\ z_2\\ z_3\end{pmatrix}=(0,0,-3)\begin{pmatrix}z_1\\ z_2\\ z_3\end{pmatrix}=-3z_3$$

$c_0=4$

(5.3)

$$c^TA^{\#}y=(0,2,0)\begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix}=(0,2)\begin{pmatrix}y_1\\ y_2\end{pmatrix}=2y_2$$

$$d^TPz=(0,2,0)\begin{bmatrix}0 & 0 & 1\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{pmatrix}z_1\\ z_2\\ z_3\end{pmatrix}=(0,0,0,)\begin{pmatrix}z_1\\ z_2\\ z_3\end{pmatrix}=0$$

$$d_0=0.$$

Evidently in this case $d^TP=0$, as witness the next to the last expression. On the other hand, $c^TP\neq 0$, as witness $c^TPz=-3z_3$ in the second expression for (5.3). Hence condition (a) of the preceding section obtains and we have $R\to\infty$ even though

$$-1\leqslant y_1\leqslant 2$$

(5.4)

$$1\leqslant y_2\leqslant 5.$$

This occurs because z is arbitrary and can be freely chosen in

$$\max R(y,z)\equiv\frac{c^TA^{\#}y+c^TPz+c_0}{d^TA^{\#}y+d^TPz+d_0}=\frac{3y_1-3z_3+4}{2y_2}, \tag{5.5}$$

*It may be observed that the $A^{\#}$ choice is not unique.

which is the specialization of (1.5) to this case. Of course, the result $R \to \infty$ in (5.1) can be confirmed by direct inspection, since negative values of z_3 may be selected along with increasingly positive values of x_1 as required in order to maintain the first interval programming constraint.

This last remark suggests that an adjunction such as

$$0 \leq x_3 \leq 1 \tag{5.6}$$

will convert (5.1) to a problem with a finite maximum. This yields an A for an interval programming format with the full row rank condition fulfilled as in

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad A^{\#} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.7}$$

On the other hand, A is also of full column rank so that we also have $A^{\#} = A^{-1}$ and

$$P = I - A^{\#}A = 0.$$

Hence both of the conditions specified in (1.6) are fulfilled, viz.

$$c^T P = d^T P = 0$$

for any c^T and d^T.

The problem to be solved is now written

$$\max R(y) \equiv \frac{c^T A^{\#} y + c_0}{d^T A^{\#} y + d_0} = \frac{y_1 - 4y_3 + 4}{2y_2}$$

(5.8) with

$$-1 \leq y_1 \leq 2$$

$$1 \leq y_2 \leq 5$$

$$0 \leq y_3 \leq 1.$$

Evidently the solution to this problem is $y_1^* = 2$, $y_3^* = 0$, and $y_2^* = 1$ so that

$$\max R(y) = \frac{10}{2} = 5.$$

To obtain the corresponding components of x we simply utilize (1.4) with $P = 0$ to obtain

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = A^{\#} y = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}. \tag{5.9}$$

As may be seen, these x values satisfy (5.1) with (5.6) adjoined. They are evidently also maximal with $R(x) = 5$ since x_3 can no longer be negative and x_2 and x_1 are at their lower and upper limits, respectively.

In some cases the solutions, as above, may be obvious but, of course, this cannot always be expected. Recourse to the preceding development, however, will produce the wanted results in any case, however, as we illustrate by now developing the above examples, along with the related background materials, in some detail as follows:

Because the denominator is unisignant we utilize section III. Observing that $d_j = d_2 = 2$ in the denominator and hence is nonnegative, we have recourse only to the first part of (3.1) in order to write

$$y_1 - a_1 = g_1\xi_1$$

6.1) $$y_2 - a_2 = g_2\xi_2$$

$$y_3 - a_3 = g_3\xi_3,$$

where, respectively, $a_1 = -1$, $a_2 = 1$, and $a_3 = 0$, via (5.8). The development from (3.1) to (3.2) applies to this case as,

with

$$0 \leqslant \xi_1 \leqslant \delta_1 = \frac{b_1 - a_1}{g_1} = \frac{2 - (-1)}{g_1}$$

(6.2) $$0 \leqslant \xi_2 \leqslant \delta_2 = \frac{b_2 - a_2}{g_2} = \frac{5 - 1}{g_2}$$

$$0 \leqslant \xi_3 \leqslant \delta_3 = \frac{b_3 - a_3}{g_3} = \frac{1 - 0}{g_3}\,.$$

The insertion of (6.1) into the functional then produces

(6.3) $$\frac{3(g_1\xi_1 + a_1) - 4(g_3\xi_3 + a_3) + 4}{2(g_2\xi_2 + a_2)} = \frac{3g_1\xi_1 - 4g\xi_3 + 1}{2(g_2\xi_2 + 1)}$$

via (5.8). Choosing

(6.4) $$g_1 = 1/2, \qquad g_2 = 1, \qquad g_3 = 1/2$$

and setting

(6.5) $$\gamma_1 = 3/2, \qquad \gamma_2 = 0, \qquad \gamma_3 = -4/2, \qquad \gamma_0 = 1/2$$

gives the denominator form wanted for (3.3) as:

$$\max R(\xi) = \frac{\frac{3}{2}\xi_1 - \frac{4}{2}\xi_3 + \frac{1}{2}}{\xi_2 + 1}$$

with

$$0 \leqslant \xi_1 \leqslant 6$$

(6.6)

$$0 \leqslant \xi_2 \leqslant 4$$

$$0 \leqslant \xi_3 \leqslant 2.$$

The transformation $\xi_j = \eta_j/\eta_0$ from our previously developed theorem [8] then produces the following example for (3.4):

$$\max \frac{3}{2}\eta_1 + 0\eta_2 - 2\eta_3 + \frac{\eta_0}{2}$$

with

$$\eta_1 + \eta_2 + \eta_3 + \eta_0 = 1$$

(6.7)

$$\eta_1 \qquad -6\eta_0 \leqslant 0$$

$$\eta_2 \quad -4\eta_0 \leqslant 0$$

$$\eta_3 - 2\eta_0 \leqslant 0$$

$$\eta_1, \eta_2, \eta_3 \geqslant 0,$$

where the g_j values of (6.4) combine with (6.2) to give $\delta_1 = 6$, $\delta_2 = 4$ and $\delta_3 = 2$, as required for the application of (3.4). The corresponding dual, which our previous theory also gives access to, is

$$\min u$$

with

$$u + \omega_1 \qquad \geqslant \frac{3}{2}$$

(6.8)

$$u + \quad \omega_2 \quad \geqslant 0$$

$$u + \qquad \omega_3 \geqslant -2$$

$$u - 6\omega_1 - 4\omega_2 - 2\omega_3 = \frac{1}{2}$$

$$\omega_1, \omega_2, \omega_3 \geqslant 0.$$

This, of course, is the application of (3.5) to the present example.

Since $\gamma_1 = 3/2$ exceeds $\gamma_0 = 1/2$ we are in the situation of case (ii) following (3.5). Thus, preserving our subscript identifications from (6.5), we have

$$\gamma_1 > \gamma_0 \geqq \gamma_2 \geqq \gamma_3 \tag{6.9}$$

in our present situation. We therefore see that the first application of the suggested algorithm should suffice. (See the remarks which conclude the case $\hat{D} > 0$ in section III.)

Applying (3.11) now produces

$$u^1 = (1/2 + 3/2 \cdot 6)/(1+6) = 19/14 > \gamma_0 = 1/2. \tag{6.10}$$

Evidently this also formally satisfies the condition that u^1 equals or exceeds the immediate successor of γ_1 is (6.9). Hence, we have

$$\begin{aligned} u^1 &= u^* = 19/14 \\ \omega_1^1 &= \omega_1^* = \gamma_1 - u^* = 3/2 - 19/14 = 2/14 \\ \omega_2^1 &= \omega_2^* = 0 \\ \omega_3^1 &= \omega_3^* = 0, \end{aligned} \tag{6.11}$$

which satisfy the constraints of (6.8), as may be verified, with $\min u = u^* = 19/14$.

Moving to the primal problem via (3.9),

$$\begin{aligned} \eta_0^1 &= \frac{1}{1+\delta_1} = \frac{1}{1+6} = \frac{1}{7} \\ \eta_1^1 &= \frac{\delta_1}{1+\delta_1} = \frac{6}{1+6} = \frac{6}{7} \end{aligned} \tag{6.1}$$

and all other $\eta_j^1 = 0$. See (3.10).

Inserting these values for the corresponding η_j in (6.7), we see that all constraints are satisfied with

$$3/2\,\eta_1 + \eta_0/2 = 3/2 \cdot 6/7 + 1/7 \cdot 1/2 = 19/14,$$

the same as the value of u^*, thereby confirming optimality. In fact, as our theory [8] prescribes, we need merely apply the expressions

$$\eta_0 = 1/7, \xi_1 = \eta_1/\eta_0 = 6/7 \div 1/7 = 6$$

with all other $\eta_j = 0$ and then reverse the development from (6.6) to (6.7) in order to verify that this value is also optimal for

$$\max R(\xi) = \frac{3/2 \cdot 6 + 1/2}{6+1} = 19/14$$

with

$$0 \leqslant \xi_1 = 6 \leqslant \frac{2-(-1)}{g_1} = 6$$

(6.13)

$$0 \leqslant \xi_2 = 0 \leqslant \frac{5-1}{g_2} = 4$$

$$0 \leqslant \xi_3 = 0 \leqslant \frac{1-0}{g_3} = 2.$$

Evidently we can now directly effect substitutions in (6.1) and obtain

$$y_1 = g_1\xi_1 + a_1 = 3 - 1 = 2$$

(6.14)

$$y_2 = g_2\xi_2 + a_2 = 0 + 1 = 1$$

$$y_3 = g_3\xi_3 + a_3 = 0 + 0 = 0.$$

Then we can proceed exactly as in (5.9) to obtain the values $x_1 = 2$, $x_2 = 1$, $x_3 = 0$ which we previously observed to be optimal.

SUMMARY

Although the development in this paper proceeded, for clarity, by algorithmic format, we summarize below the explicit solution in tabular format for direct theoretical interpretation and utilization.

Summary of Solution

Denominator	Transformed numerator	R^*
Bisignant.....		∞
Unisignant:		
(a) positive.....		$u^s = \dfrac{\gamma_0 + \sum_1^s \gamma_2\delta_2}{I + \sum_1^s \delta_2}$, least s with $u^s \geqq \gamma_{s+1}$
(b) nonnegative.....	$\gamma_0 > 0$	∞
	$\gamma_0 = 0$	γ_1
	$\gamma_0 < 0$	$u^s = \dfrac{\gamma_0 + \sum_1^s \gamma_2\delta_2}{\sum_1^s \delta_2}$, least with $u^s \geqq \gamma_{s+1}$

CONCLUSION

Before proceeding any further we should probably point up, again, the crucial role played by the general theory (including the transformations and proof procedures) which we introduced in [8] for making explicit contacts between linear fractional and ordinary linear programming, in all generality and exact detail. These transformations are also utilized in the present paper and the theory is also extended via the duality (and other) characterizations given in the preceding text. These are joined together here for the proofs in algorithmic format kind we have just illustrated by example and commentary. Other uses can undoubtedly also be made of this theory and the preceding extensions via the passage (up and back) between linear fractional and ordinary linear programming that is now possible.

Our general theory has been used by others, too, to extend or simplify parts of linear fractional programming en route to effecting the contacts with ordinary linear programming that are thereby obtained. The work of Zionts [22] should perhaps be singled out as being most immediately in line with the $R = \infty$ results presented in this paper. Zionts' development is directed only toward simplifying matters by focusing on eliminating cases for linear fractional programming which are either deemed to be unwanted or of little interest for practical applications.

The developments cease as soon as the contacts with ordinary linear programming are identified via our theory, which he like others utilizes for this purpose.

We have effected the developments in this paper in a way that makes contact with interval linear programming.* An opening for further two-way flows is thereby also provided. The resulting junctures should also help to guide subsequent developments in the more special situations that now seem to invite consideration in the future. Finally, the possibilities for dealing with specially structured problems (such as those observed at the start of the present paper) should also be observed explicitly in this conclusion, partly because the theory we have now developed and presented should also be a helpful guide to these additional cases which are important in their own right. Thus we can now conclude here by referring to our opening remarks.

ACKNOWLEDGMENT

We wish to thank W. Szwarc of the University of Wisconsin for comments which helped us to improve the exposition in the manuscript for this article.

BIBLIOGRAPHY

[1] Balas, E. and M. Padberg, "Equivalent Knapsack-Type Formulations of Bounded Integer Programs," Carnegie-Mellon University (Sept. 1970).

[2] Ben-Israel, A. and A. Charnes, "An Explicit Solution of a Special Class of Linear Programming Problems," Operations Research *16*, 1166–1175 (1968).

[3] Ben-Israel, A., A. Charnes, and P. D. Robers, "On Generalized Inverses and Interval Linear Programming," Proceedings of The Symposium on Theory and Applications of Generalized Inverses, held at Texas Technological College, Lubbock, Tex. (Mar. 1968).

[4] Ben-Israel, A. and P. D. Robers, "A Decomposition Method for Interval Linear Programming," Management Science Vol. 16, No. 5 (Jan. 1970).

[5] Bradley, G., "Transformation of Integer Programs to Knapsack Problems," Yale University, Rept. No. 37 (1970). To appear in *Discrete Mathematics*.

*See [2], [3], [4].

[6] Chadda, S. S., "A Decomposition Principle for Fractional Programming," Opsearch *4*, 123–132 (1967).

[7] Charnes, A. and W. W. Cooper, *Management Models and Industrial Applications of Linear Programming* (John Wiley & Sons, Inc., New York, 1961).

[8] Charnes, A. and W. W. Cooper, "Programming with Linear Fractional Functionals," Nav. Res. Log. Quart. *9*, 181–186 (Sept.–Dec., 1962).

[9] Dorn, W. S., "Linear Fractional Programming," IBM Research Report RC–830 (Nov. 27, 1962).

[10] Derman, C., "On Sequential Decisions and Markov Chains," Management Science, *9*, 16–24 (1962).

[11] Gilmore, P. C. and R. E. Gomory, "A Linear Programming Approach to the Cutting Stock Problem," Operations Research *11*, 863–888 (1963).

[12] Glover, F. and R. E. Woolsey, "Aggregating Diophantine Equations," University of Colorado Report 70–4 (Oct. 1970).

[13] Isbell, J. R. and W. H. Marlow, "Attrition Games," Nav. Res. Log. Quart. *3*, 71–93 (1956).

[14] Jagannathan, R., "On Some Properties of Programming in Parametric Form Pertaining to Fractional Programming," Management Science, *12*, 609–615 (1966).

[15] Joksch, H. C., "Programming with Fractional Linear Objective Function," Nav. Res. Log. Quart. *11*, 197–204 (1964).

[16] Klein, M., "Inspection-Maintenance-Replacement Schedules under Markovian Deterioration," Management Science *9*, 25–32 (1962).

[17] Martos, B., "Hyperbolic Programming," translated by A and V. Whinston, Nav. Res. Log. Quart. *11*, 135–155 (1964).

[18] Robers, P. D., "Interval Linear Programming," Ph.D. Thesis submitted to Northwestern University (Evanston, Ill.) Dept. of Industrial Engineering and Management Sciences (1968).

[19] Robers, P. D. and A. Ben-Israel, "A Suboptimization Method for Interval Linear Programming," Systems Research Memo No. 206, Northwestern University (Evanston, Ill.), The Technological Institute (June 1968).

[20] Swarup, K. "Linear Fractional Functionals Programming," Operations Research *13*, 1029–1036 (1965).

[21] Wagner, H. M. and J. S. C. Yuan, "Algorithmic Equivalence in Linear Fractional Programming," Management Science *14*, 301–306 (Jan. 1968).

[22] Zionts, S., "Programming with Linear Fractional Functionals," Nav. Res. Log. Quart. *15*, 449–452 (Sept. 1968).

USING DECOMPOSITION IN INTEGER PROGRAMMING

Linus Schrage

Graduate School of Business
University of Chicago

ABSTRACT

When implicit enumeration algorithms are used for solving integer programs, a form of primal decomposition can be used to reduce the number of solutions which must be implicitly examined. If the problem has the proper structure, then under the proper decomposition a different enumeration tree can be defined for which the number of solutions which must be implicitly examined increases with a power of the number of variables rather then exponentially. The proper structure for this kind of decomposition is that the southwest and northeast corners of the constraint matrix be zero or equivalently that the matrix be decomposable except for linking columns. Many real traveling salesmen, plant location, production scheduling, and covering problems have this structure.

INTRODUCTION

Consider an integer program for which the constraint matrix has the form shown in Figure 1. The important feature is that the columns of A_2 link two otherwise independent subproblems.

In general, a set of columns is a linking set if deleting that set partitions all remaining columns

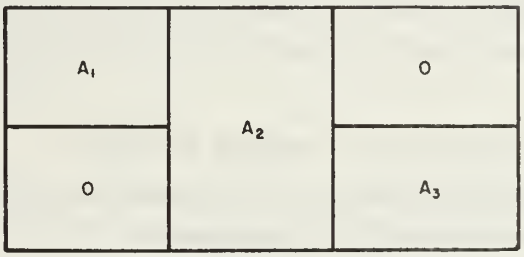

FIGURE 1.

into two disjoint, nonempty sets, A_1 and A_3, such that there do not exist two columns, one in A_1, and one in A_3, both having a nonzero entry in the same row.

For simplicity, assume that all variables are required to be either 0 or 1, and that there are n variables in each of the blocks A_1, A_2, and A_3. There are then 2^{3n} solutions to be implicitly examined.

FORM OF THE DECOMPOSITION

We assume that an enumerative scheme similar to that described in Geoffrion [4] is to be used. We can think of A_2 as being the master problem. If we fix the values of the variables in block A_2 to some set of values, then the variables in blocks A_1 and A_3 constitute independent subproblems and can be solved separately. Solving one of these problems (A_1 or A_3) requires us to implicitly examine 2^n solutions. Each of these two problems must be solved for each possible setting of the variables in A_2. Therefore, we must effectively enumerate $2^n(2^n+2^n)=2^{2n+1}$ solutions. If we define a node to be the setting of some variable to one of its two values, then perhaps a better measure of computational

difficulty is the number of nodes in the enumeration tree. The nondecomposition complete enumeration tree has approximately 2^{3n+1} nodes in it. By use of decomposition the complete enumeration tree has $2^n(2^{n+1}+2^{n+1})+2^{n+1}=2^{2n+2}+2^{n+1}$ nodes. In this sense, decomposition reduces the difficulty of the problem by a factor of approximately 2^{n-1}.

EXISTENCE OF THE DECOMPOSITION STRUCTURE

One can argue that this stairstep structure exists in several classes of integer programs. In a traveling salesman problem based on the United States, the variables in A_2 might correspond to the choice of arcs connecting cities in the Midwest. For each set of these arcs chosen, one should be able to complete the eastern and western legs of the tour independently.

Similar arguments can be given for the plant location problem based on cities in the United States. The variables in A_2 would correspond to the decisions of which plants to build in the midwest. For each set of midwest plant decisions one would expect to be able to solve the eastern and western plant location problems independently.

The Wagner-Whitin [10] dynamic lotsize problem is a special case of the plant location problem which has an even more obvious decomposition structure. Once we decide to produce in a particular period, the production plans for subsequent and for previous periods can be solved independently. By discarding variables which obviously cannot be in the optimal solution, the 12-period problem given in Wagner and Whitin [10] can be decomposed into a stairstep structure with seven blocks.

Another class of integer programs are covering problems. One situation giving rise to covering problems is in the assignment of vehicles to routes. Each city in the service area must be "covered" by at least one route. Arguments similar to those for the traveling salesman and plant location problems can then be made.

GENERALIZATION TO MORE THAN THREE BLOCKS

Consider a constraint matrix with the stairstep structure shown in Figure 2. Again, for simplicity, we assume the problem is composed only of 0–1 variables.

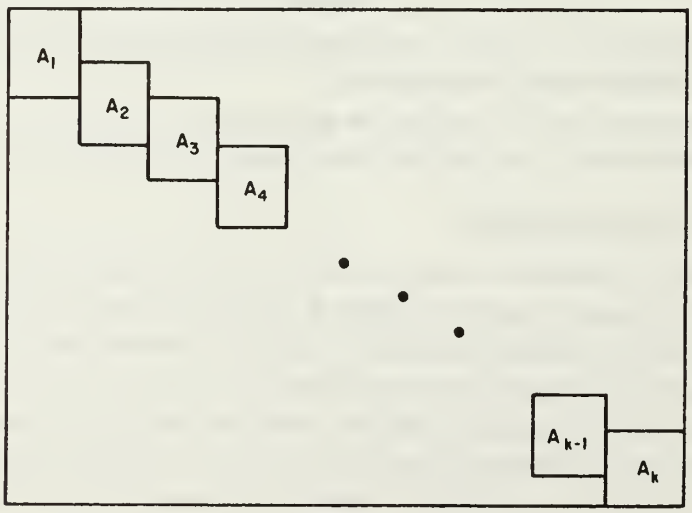

FIGURE 2.

This problem can be decomposed into a hierarchy of masters and subproblems. Assume there is a total of k blocks. Choose as the highest level master, block number $[k/2]+1$, where $[x]$ is the greatest

integer no larger than x. This divides the problem into two independent subproblems with $[k/2]$ and $[(k-1)/2]$ blocks each. These two subproblems can themselves be decomposed in similar fashion. If this decomposition is carried on in this recursive fashion we will then have approximately $\log_2 k$ levels of decomposition. In doing the enumeration we will first set the variables in block $[k/2]+1$. For each setting of these variables we must solve the independent subproblems composed of blocks 1 to $[k/2]$ and blocks $[k/2]+2$ to k. The first of these subproblems is solved by first setting the variables in block $[[k/2]/2]+1$. This divides the problem further into independent subproblems. This enumeration method is applied recursively to each independent subproblem created. The result is a binary tree of independent subproblems.

For simplicity assume that each block contains n 0–1 variables. Let $s(j)$ be the number of terminal nodes in a decomposition enumeration tree with j levels. If we increase the number of blocks such that the number of levels of decomposition increases from j to $j+1$, then each of the 2^n solutions to the highest level master partitions the remainder of the problem into two j level problems. We then have that $s(j+1)=2^n \cdot 2 \cdot s(j)$, where $s(1)=2^n$. Thus, $s(j)=(2^{n+1})^{j/2}$ and the number of terminal nodes in the decomposition tree of a problem with k blocks is $(1/2)(2^{n+1})^{\log_2 k}=(1/2)k^{n+1}$. A perhaps more accurate measure of the size of the enumeration tree is the total number of nodes, terminal and intermediate, in the tree. Let $t(j)$ be the total number of nodes in the decomposition enumeration tree with j levels of decomposition. The total number of nodes in a simple binary tree with 2^n terminal nodes is $2^{n+1}-1$ or approximately 2^{n+1}. We can now argue as before the claim that approximately $t(j+1)=2^n \cdot 2 \cdot t(j) + 2^{n+1} = 2^{n+1}(t(j)+1)$. Now, $t(1)$ is approximately 2^{n+1}. Thus $t(j)$ is approximately

$$\sum_{i=1}^{j} (2^{n+1})^i = [2^{n+1} - (2^{n+1})^{j+1}]/[1-2^{n+1}]$$

which is approximately $(2^{n+1})^j$ for n and j large. Thus, the total number of nodes in the decomposition tree of a problem with k blocks is approximately $(2^{n+1})^{\log_2 k} = k^{n+1}$. We now see that the number of solutions and the total number of nodes which must be implicitly examined increases with a power of the number of variables if the block size remains constant as the problem size is increased. The number of solutions which would have to be implicitly examined under no decomposition is 2^{kn} while the number of nodes in the tree under no decomposition is $2^{kn+1}-1$. For example, if $n=10$ and $k=3$, then decomposition decreases the size of the tree by a factor of about 500. If $k=7$ the factor is about 10^9.

Edmonds [3] made the interesting suggestion that a problem be considered tractable if and only if one can exhibit an algorithm for its solution whose running time is bounded by a polynomial in the size of problem. There is no known algorithm for general integer programs which is polynomial bounded. In fact, Karp [5] gives "theorems which strongly suggest, but do not imply, that these problems, as well as many others, will remain intractable perpetually." It is interesting therefore that the class of decomposable integer programs just described can be solved by a polynomial bounded algorithm, namely, simply searching the decomposition enumeration tree.

The Wagner-Whitin [10] example problem, for example, is a 0–1 problem with 12 variables. The first variable is required to be 1 so there are $2^{11}=2{,}048$ feasible solutions to the problem and 1,095 nodes in the simple enumeration tree. Using the seven block decomposition mentioned earlier, the equivalent of only 96 solutions need be examined implicitly. The number of nodes in this decomposition tree is 191. The computation involved in an enumerative algorithm should be approximately

proportional to the number of nodes examined. Complete enumeration of the 191-node decomposition tree in this case is not an unreasonable solution method.

MORE THAN TWO BLOCKS PER DECOMPOSITION

In the analysis thus far we have assumed that a linking block decomposed a problem into two independent subproblems. Much the same analysis could be done if a set of linking columns decomposed a problem into more than two independent subproblems. See, for example, Figure 3.

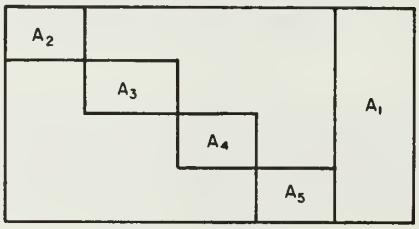

FIGURE 3.

For each of the settings of the variables in A_1 which must be examined, the subproblems A_2, A_3, A_4, and A_5 can be solved independently.

PROBLEMS WITH LINKING CONSTRAINTS

The more common form of decomposition structure discussed in linear programming literature involves a set of submatrices with no rows in common, except that there is a set of constraints at the top linking all the submatrices together. It may be that problems which are formulated with that structure actually have a structure like Figure 2. The rows in common between A_1 and A_2, A_3 and A_4, etc. could be moved to the top and then one would have the more common form of decomposition structure.

The decomposition approach described may also be useful for problems where the natural formulation is with linking constraints by realizing that a linking constraint can be replaced by a linking column and two nonlinking constraints. Consider the linking constraint:

$$\sum_{j=1}^{2n} a_j x_j = b.$$

Suppose that the problem is decomposable into two subproblems, each with n variables, except for this constraint. This linking constraint can be replaced by one linking variable, y_1, and two nonlinking constraints as follows:

$$\sum_{j=1}^{n} a_j x_j - y = 0$$

$$y + \sum_{j=n+1}^{2n} a_j x_y = b.$$

Assume again that the x_j must be either 0 or 1 and all the a_j are integer. In the worst possible case we must examine 2^n different values for y, and the size of the tree under decomposition actually doubles.

We would expect the number of different values of y to be examined to be much less than 2^n. In a covering problem, for example, the a_j's are either 0 or 1. Then, in the worst possible case we must examine n different values of y. The number of terminal nodes in the decomposition tree is then $n(2^n+2^n) = n2^{n+1}$ versus 2^{2n} terminal nodes in the nondecomposition tree. If $n=20$ for example, then the number of terminal nodes is reduced by a factor of approximately 26,000.

If one considers the usual decomposition structure studied in linear programming where there are p independent subproblems row-linked by a single master block with q rows, then the interested reader should be able to convince himself that the proper generalization of the transformation described in the previous problem is to reformulate the problem as a multilevel decomposition problem with $p-1$ sets of linking columns and approximately $\log_2 p$ levels of decomposition. Each linking set would consist of q columns.

MIXED VARIABLES CASE

We have considered only pure 0–1 integer programs thus far. The analysis generalizes fairly naturally to the case where some of the variables may take on any integer value. The analysis also extends to the case where some of the variables are not required to be integral if fixing the integer variables in a block completely implies values for the continuous variables in that block.

DISADVANTAGES OF THIS DECOMPOSITION METHOD AND THEIR PARTIAL ALLEVIATION

A disadvantage of this decomposition method is that the order in which variables are placed in the enumeration tree is partially specified beforehand. The importance of flexibility in the tree search, especially in specifying the order of addition of variables to the enumeration tree has been pointed out [2], [4], [6], [7], [9]. For example, if there is a variable which can take on the value of either 0 or 1 in the optimal solution (i.e., there are alternate optima), then the amount of searching required by a branch-and-bound or implicit enumeration algorithm is approximately doubled if this variable is placed first in the tree rather than last. If there are unimportant variables in the highest level master problem, the performance of an implicit enumeration aldorithm could be appreciably degraded by the decomposition approach.

Another apparent disadvantage is the additional bookkeeping which must be done to keep track of the decomposition tree.

The first disadvantage can be alleviated somewhat by adapting the flexible tree-search procedure described in Tuan [9] and Bravo, et al. [2]. This approach allows one to partially reorder a tree without re-searching branches already searched or discarding branches yet to be searched. With respect to the second disadvantage, the bookkeeping scheme is not significantly more complex than conventional ones.

The additional restrictions on the branching scheme are that

(a) we cannot branch on a variable i unless all variables in the master of the subproblem containing i have been fixed;

(b) we can backtrack any time the bound on the subproblem currently being solved is worse than the value of some feasible solution to the subproblem for the current setting of the variables in the master problem; or

(c) we can backtrack any time that the overall bound is worse than the value of some known feasible solution to the entire problem.

COMPUTATIONAL EVALUATION

A computer program was written incorporating the decomposition method into a backtrack implicit enumeration scheme. The program was similar to those described by Geoffrion [4]. The revised simplex method of linear programming with explicit inverse was used to calculate bounds. After any variable was forced to 0 or 1, dual pivots were performed to return to feasibility. After any variable in a backtrack step was released from 0 or 1, primal pivots were performed to return to optimality. The rule for selecting the next variable to force to 0 or 1 was simply to force to the nearer integer value that basic variable which was closest in value to an integer.

The implementation was inefficient in at least two ways: 1) the *LP* portion worked with a full inverse at all times. That is, even though only a subproblem was being optimized, pivots were done in the full inverse. For the problems considered, each subproblem had half as many rows as the full problem, therefore each pivot in the implemented procedure may have taken as much as four times as much work as really necessary. 2) Natural integrality in the master problem was not taken advantage of. Before a subproblem can be searched, each variable in its master must be fixed to an integer value. If some variable x_j in the master was fixed to 1 and would have remained at the value 1, even if not constrained to the value 1, all through the enumeration of all subproblems, then it would not be necessary to examine $x_j=0$ in the backtrack step. Most integer programming routines take advantage of this natural integrality. The program here did not.

A class of decomposable integer programs was derived based on a problem known as IBM–1. A description of this problem can be found in Trauth and Woolsey [8]. The problem is a general integer program with seven variables and seven constraints. Two 0–1 variables were required to represent each of the original seven variables. A series of eight problems with the stairstep structure shown in Figure 2 was created. Each problem was composed of three blocks. Each of the two outer blocks consisted of a copy of the IBM–1 constraint matrix. In eight different problems, the middle linking block consisted of the first k columns of IBM–1, where k ranged from zero to seven. All problems had 14 rows.

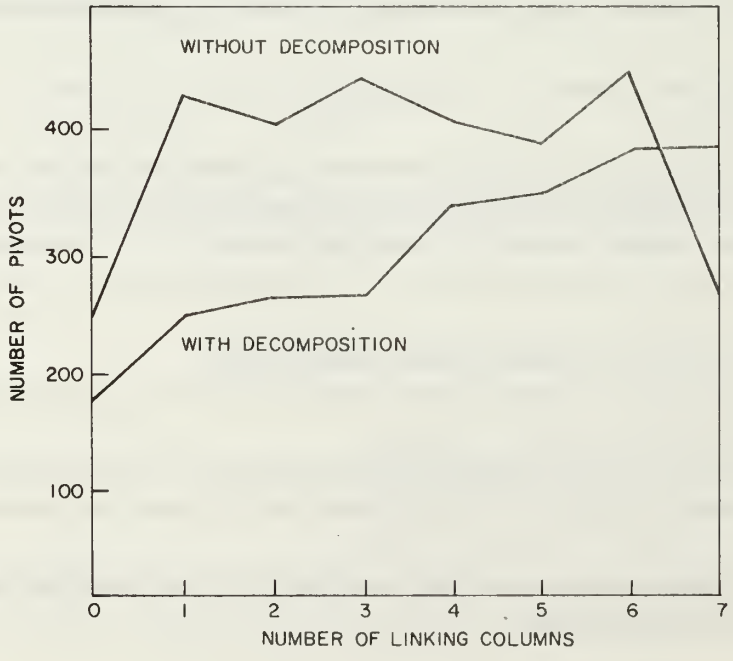

FIGURE 4.

These problems were solved on an IBM 360/65. This machine uses multiprogramming; thus run times are random variables. The number of pivots is therefore perhaps a more reliable estimate of computational difficulty because most of the work is involved in pivoting. This statistic is plotted versus number of linking columns in Figure 4. As expected, the advantage of decomposition tends to diminish as the number of linking columns increases. The same program was used to solve the problems without decomposition; the program was simply not told that the problem was decomposable. Recall that under decomposition, each pivot should require less work than under no decomposition.

Other computational statistics are displayed in Table 1. The link-edit time column is included only to give an indication of the variability in run times. The link-edit step required exactly the same amount of work in every run.

TABLE 1. *A Comparison of Decomposition with No Decomposition*

No. of linking columns	Pivots		Time (sec)		Nodes examined		Link-edit time (sec)	
	With decomp.	With-out	With decomp.	With-out	With decomp.	With-out	With decomp.	With-out
0	174	246	4.50	6.22	48	79	2.62	2.80
1	241	427	6.07	7.45	52	106	2.55	2.75
2	257	404	8.12	8.07	57	108	3.15	2.70
3	255	445	7.09	8.40	60	109	2.87	2.74
4	334	406	8.84	7.89	68	102	3.24	2.60
5	346	383	8.92	7.39	70	78	2.55	2.67
6	384	446	9.97	9.07	87	73	2.47	2.85
7	384	262	10.40	9.29	99	86	2.84	2.92

The run times under decomposition are perhaps longer than they need be in practice because a full solution report was printed each time a better integer solution was obtained to any subproblem. The runs without decomposition would typically produce three solution reports while a run under decomposition would typically produce, say, nine solution reports. A full solution report requires a fair amount of work because the inverse must be multiplied through the full matrix to calculate such things as reduced costs and dual prices.

The decomposition method for these problems seems to be fairly robust in that the amount of work seems to increase less than linearly with number of linking columns for this class of problems. For a small number of linking columns, decomposition is clearly superior.

REFERENCES

[1] Balas, E., "An Additive Algorithm for Solving Linear Programs with Zero-One Variables," Operations Research *13*, 517–546 (1965).

[2] Bravo, A., J. G. Gomez, L. Lustosa, L. Schrage, and N. Pizzolato, "A Mixed Integer Programming Code," CMSBE Report No. 7043, University of Chicago (Sep. 1970).

[3] Edmonds, J., "Paths, Trees, and Flowers," Canadian J. Math. *27*, 449–467 (1965).

[4] Geoffrion, A. M., "An Improved Implicit Enumeration Approach for Integer Programming," Operations Research *17*, 437–454 (May-June 1969).

[5] Karp, R. M., "Reducibility Among Combinatorial Problems," presented at ORSA National Convention, New Orleans, La. (Apr. 26, 1972).

[6] Salkin, H. M., "On the Merit of Generalized Origin and Restarts in Implicit Enumeration," Operations Research *18*, 549–555 (May-June 1970).

[7] Spielberg, K., "Plant Location with Generalized Search Origin," Management Science *16*, 165–178 (1969).

[8] Trauth, C. A. and R. E. Woolsey, "Integer Linear Programming: A Study in Computational Efficiency," Management Science *15*, 481–493 (May 1969).

[9] Tuan, Nghiem Ph., "A Flexible Tree-Search Method for Integer Programming Problems," Operations Research *19*, 115–119 (Jan.-Feb. 1971).

[10] Wagner, H. and T. M. Whitin, "Dynamic Version of the Economic Lot Size Model," Management Science Vol. 5, No. 1 (Oct. 1958).

NUMERICAL TREATMENT OF A CLASS OF SEMI-INFINITE PROGRAMMING PROBLEMS*

S. Å. Gustafson

The Royal Institute of Technology

Stockholm, Sweden

and

K. O. Kortanek

Carnegie-Mellon University

Pittsburgh, Pennsylvania

ABSTRACT

Many optimization problems occur in both theory and practice when one has to optimize an objective function while an infinite number of constraints must be satisfied. The aim of this paper is to describe methods of handling such problems numerically in an effective manner. We also indicate a number of applications.

1. INTRODUCTION

In order to illustrate the subject of this paper, we immediately give a few examples on the class of problems we wish to study.

EXAMPLE 1.1: One wants to determine a cumulative distribution function G which corresponds to a stochastic variable, which can assume values inside a finite closed interval $[a, b]$. In N points $t_1, t_2, \ldots, t_N$, one has measured the values of G and obtained $g_1, g_2, \ldots, g_N$. We want to approximate G in $[a, b]$ by a polynomial P of a degree less than a certain number n. It is natural to write

$$P(t) = \sum_{r=1}^{n} y_r t^{r-1}$$

and require that $P(a) = 0$, $P(b) = 1$, and $P'(t) \geq 0$. Since we cannot hope that P passes through the measured points, we try to solve the problem

$$\inf_{y_1, y_2, \ldots, y_n} \sum_{j=1}^{N} \left(\sum_{r=1}^{n} y_r t_j^{r-1} - g_j \right)^2,$$

subject to

$$\sum_{r=1}^{n} y_r a^{r-1} = 0$$

$$\sum_{r=2}^{n} (r-1) y_r t^{r-2} \geq 0 \qquad a \leq t \leq b$$

$$\sum_{r=1}^{n} y_r b^{r-1} = 1.$$

*This research was supported in part by National Science Foundation Grant GK 31-833.

It is easily shown by examples that this problem may or may not be feasible, depending on the given data.

This problem appeared when one wanted to study size distributions of grains in gravel deposits in order to get a suitable raw material for concrete production (Gustafson-Martna [27]). In the referenced paper, one did not attempt to solve exactly the problem indicated above, but used piecewise polynomial interpolation through the measured points instead, in order to meet the monotonicity condition.

EXAMPLE 1.2: Bojanic-DeVore [4] discuss the problem of one-sided approximation of a given function from below, while maximizing a linear functional. Their problem can be stated as follows: Let $[a, b]$ be a closed interval, $u_1, u_2, \ldots, u_n$ n given functions which form a Čebyšev system (for a definition see, e.g., Karlin-Studden [32] or Gustafson [20]). Let further ϕ be continuous on $[a, b]$. Determine

$$\max \sum_{r=1}^{n} y_r \int_a^b u_r(t)dt,$$

subject to

$$\sum_{r=1}^{n} y_r u_r(t) \leq \phi(t), \qquad t \in [a, b].$$

Bojanic-DeVore give some unicity and existence results and identify the solutions with certain quadrature rules. Methods for finding computational solutions to this problem are given in Gustafson [20].

EXAMPLE 1.3: Let again $[a, b]$ be a closed interval, ω a positive function defined on $[a, b]$ and ϕ continuous on the same interval. We want to determine a polynomial of degree less than n which approximates ϕ as well as possible in the weighted maximum norm determined by ω. That is, we want to solve the problem.

Compute min η subject to

$$\omega(t)\left|\sum_{r=1}^{n} y_r t^{r-1} - \phi(t)\right| \leq \eta \qquad t \in [a, b].$$

We can write this task in the equivalent form:

Compute min η subject to

$$\sum_{r=1}^{n} y_r t^{r-1}\omega(t) - \eta \leq \phi(t)\omega(t) \text{ and } -\sum_{r=1}^{n} y_r t^{r-1}\omega(t) - \eta \leq -\phi(t)\omega(t).$$

This problem is well known and for $\omega(t) = 1$, a solution is constructed computationally by means of Remez' algorithms (see, e.g., Cheney [9]). For numerical purposes, an approximative solution often is satisfactory (Powell [41]), Gustafson-Dahlquist [23]).

EXAMPLE 1.4: Kantorovich-Rubinshtein [31] and Rubinshtein [43] give examples of production scheduling problems, where an infinite number of linear constraints must be met. Also Vershik-Temel't [48] propose a process of finding a sequence of approximate finite linear programming problems whose optimal values converge to the optimal value of the infinite linear programming problem.

EXAMPLE 1.5: Gorr-Kortanek [18], Gustafson-Kortanek [24] and Gustafson-Kortanek [25] give examples of models for study of air pollution problems, where an infinite number of linear constraints must be fulfilled over a two-dimensional set S.

Additional Examples and Problems

As illustrated above, moment problems stem from problems in approximation and minimization, see Shohat-Tamarkin [45], Rivlin-Shapiro [42], Shapiro [44], Karlin-Studden [32] and others. This leads to applications of infinite programming techniques to analysis, Duffin [13], [14], Kretschmer [36], [37], Duffin-Karlovitz [15], including the development of orthogonality theorems and similar results with application to the theory of integral equations. While applications of the moment problem to statistics and probability theory are well-known, recent problems in these areas have been brought into contact with the theory of moments by Krafft [35]. Interesting applications of generalized moment problems have also been made in theoretical physics, see Baker-Gammel [1]. See also the classification theory of Ben Israel-Charnes-Kortanek [3] for linear programming problems over closed convex sets in locally convex spaces and applications to approximation theory.

DEFINITION 1.1: We denote by *problem D* the general task:

Compute

$$\inf_{y_1, y_2, \ldots, y_n} G(y_1, y_2, \ldots, y_n),$$

subject to

$$\sum_{r=1}^{n} y_r u_r(x) \geq \phi(x) \qquad x \epsilon S. \tag{1.1}$$

Here S is a given set, $u_1, u_2, \ldots, u_n, \phi$ are given functions defined over S. The objective function G is also given and must be defined for all vectors $y_1, y_2, \ldots, y_n$ which satisfy (1.1).

We will refer to problem D as a semi-infinite program. The fact that n is finite is crucial in our analysis. We observe that all the examples, 1.1 through 1.5, are instances of problem D.

Many well-known optimization problems are subsumed by problem D. If S has a finite number of elements, we arrive at mathematical programming tasks of various kinds. Note in particular that if G also is linear, we get linear programs. For a discussion of these problems, the reader is referred to the textbooks by Charnes-Cooper [6] and Dantzig [10].

We want, instead, to discuss the case when S has an infinite number of elements. The general theory of semiinfinite programming embracing such problems is given in several papers by Charnes, Cooper, and Kortanek [7], [8]. Included in their theory is the development of regularization techniques, analogous to those of finite linear programming, which we use in our computational developments.

In section 2 of this paper we present the parts of the general theory which are relevant for our purpose. We discuss the intimate connection between problem D and certain so-called moment prob-

lems. We establish that the solutions of problem D can be found if one can solve a system of a finite number of scalar equations in a finite number of unknowns. This system is nonlinear even if G is linear and its numerical solution is a nontrivial task. In section 3 we propose an algorithm to be used in practical computational work. The basic idea underlying our algorithm is that D is approximated by a problem with a finite number of constraints. The solution hereby obtained is then used as an initial approximation which is then improved by Newton-Raphson iterations (other iterative methods might be considered). Thus the solution of semi-infinite programs can be achieved by combining well-known standard techniques. In particular problems, special short cuts can be used in order to facilitate the computations (see, e.g., Gustafson [20], [21]). We also discuss questions in connection with error estimation. We treat the problem of assessing how perturbations of input data influence the optimal solution and corresponding value of the objective function.

2. GENERAL RESULTS

DEFINITION 2.1: Let K be the set of vectors $y=(y_1, y_2, \ldots, y_n)$ which satisfy (1.1). We refer to K as the *constraint set* of D.

We note that K always is contained in R^n, the n-dimensional vector space (independent of the nature of S). If K is nonempty, it is convex. We give three simple instances of D in order to illustrate different situations that can occur.

EXAMPLE 2.1: Find

$$\inf y_1+y_2,$$

when

$$y_1x+y_2x^2 \geqslant 1, \qquad x\epsilon[0, 1]$$

(inconsistent).

EXAMPLE 2.2: Find

$$\inf y_1,$$

when

$$y_1+y_2x \geqslant \sqrt{x}, \qquad x\epsilon[0, 1]$$

(the inf.-value is 0, but it is not attained).

EXAMPLE 2.3: Find

$$\inf y_1+1/2\ y_2,$$

when

$$y_1 + y_2 x \geq \frac{1}{1+x}, \qquad x \in [0,1]$$

(The inf.-value is 3/4, and is assumed for $y_1 = 1$, $y_2 = -1/2$).

Another example is treated in lemma 3.5 in this paper. It is quite obvious that problems such as examples 2.1 and 2.2 above will cause difficulties in actual machine computaton. We therefore want to specialize problem D, but in such a way that we still retain wide generality and consider only what we call regularized problems. (Compare Gustafson [20].)

DEFINITION 2.2: We denote by *problem* D_F a special case of problem D having the properties 1,2,3 below:

1. S can be written as $S_k \cup S_F$, where S_k is a compact subset of the k-dimensional vector space ($k < \infty$) and S_F has a finite number of elements. The conditions

$$\sum_{r=1}^{n} y_r u_r(x) \geq \phi(x), \qquad x \in S_F, \tag{2.1}$$

must be such that they alone restrict y to a bounded region K_F of R^n. We require, however, that (2.1) is consistent.

2. We require further that $u_1, u_2, \ldots, u_n$ and ϕ are continuous over S_k and that $u_1, u_2, \ldots, u_n$ meet Krein's condition: There exist constants $c_1, c_2, \ldots, c_n$ such that

$$\sum_{r=1}^{n} c_r u_r(x) > 0, \qquad x \in S_k.$$

3. G must be differentiable and convex on K.

We note that for a task of type D_F, K is a compact bounded set. Hence the minimum value Z is always assumed.

Instead of Krein's condition, Gustafson [20] requires that $u_1, u_2, \ldots, u_n$ form a Čebyšev system (is a unisolvent set). Unfortunately, this cannot be done if $k > 1$. We quote the following classical result from Buck [5] (the notations are slightly changed; $C(\Omega)$ denotes the space of functions, continuous over Ω).

LEMMA 2.1: If Ω is a compact connected set and $C(\Omega)$ contains a unisolvent linear subspace of finite dimension at least 2, then Ω is homeomorphic to the unit interval or the unit circumference.

Therefore, many results in this section will be generalizations of those in Gustafson-Kortanek-Rom [26].

In order not to get unnecessarily complicated formulae, we treat only the case that the inequalities corresponding to $x \in S_F$ are of the type (2.2) below. We want first to treat the case

$$G(y_1, y_2, \ldots, y_n) = \sum_{r=1}^{n} \mu_r y_r,$$

and then generalize the results to a general D_F-problem. Denote the problem by D_{F_0}.

Compute $$Z = \min \sum_{r=1}^{n} y_r \mu_r.$$

subject to $$\sum_{r=1}^{n} y_r u_r(x) \geq \phi(x), \qquad x \epsilon S_k \text{ and}$$

(2.2) $$F_1 \leq y_r \leq F_2 \qquad r = 1, 2, \ldots, n$$

We want to show that the optimal solution of D_{F_0} can be found by solving a nonlinear system of equations. In order to derive this, we apply the theory of semi-infinite programming. For this purpose we need a few notations.

Let Σ be the set of all finite regular measures which meet the integrability conditions

$$\int_{S_k} |u_r(x)| d\alpha(x) < \infty \qquad r = 1, 2, \ldots, n$$

$$\int_{S_k} |\phi(x)| d\alpha(x) < \infty.$$

Denote by M_n the convex cone in R^n:

$$M_n = \left\{ \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \mid \sigma_r = \int_{S_k} u_r(x) d\alpha(x), \qquad r = 1, 2, \ldots, n, \qquad \sigma \epsilon \sum \right\}.$$

Introduce now the problems P_0 and D_0:

Compute $\sup_{\alpha} \int_{S_k} \phi(x) d\alpha(x)$	Compute $\inf_{y_1, \ldots, y_n} \sum_{r=1}^{n} y_r \mu_r.$
subject to $\int_{S_k} u_r(x) d\alpha(x) = \mu_r \qquad r = 1, 2, \ldots, n,$	subject to $\sum_{r=1}^{n} y_r \mu_n(x) \geq \phi(x) \; x \epsilon S_k.$

$$\alpha \epsilon \sum.$$

From Karlin-Studden [32, p. 472], we quote the result (the notations are slightly changed).

LEMMA 2.2: Let $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ belong to the interior of M_n. Then the optimal values of P_0 and D_0 are equal.

Arguing as in Gustafson [20], we can associate with D_{F_0} the semi-infinite dual problem P_{F_0}:

Compute $$\max \int_{S_k} \phi(x) d\alpha(x) + \sum_{r=1}^{n} (F_1 v_r^+ - F_2 v_r^-),$$

subject to $$\int_{S_k} u_r(x) d\alpha(x) + v_r^+ - v_r^- = \mu_r \qquad r = 1, 2, \ldots, n$$

$$\alpha \epsilon \sum \; v_n^+ \geq 0, \; v_r^- \geq 0 \qquad r = 1, 2, \ldots, n.$$

P_{F_0} is always feasible, but may be unbounded. Exactly as in Gustafson [20] and Charnes-Cooper-Kortanek [8], we establish:

LEMMA 2.3: Let D_{F_0} have interior points. Then P_F and D_F are consistent and bounded. They assume their optimal values which are equal.

We also obtain, following [8], [20]:

LEMMA 2.4: Among the optimal solutions of P_{F_0}, there are such which correspond to point-masses with finite number of mass-points. Furthermore,

LEMMA 2.5: Let D_{F_0} have interior points and let an optimal solution of P_{F_0} be given by

i) a pointmass distribution with mass m_i at x^i, $i=1, 2, \ldots, q$,

ii) q^+ of the v_r^+ are positive, namely $v_{r_1}^+, v_{r_2}^+, \ldots, v_{r_{q^+}}^+$,

iii) q^- of the v_r^- are positive, namely $v_{s_1}^-, v_{s_2}^-, \ldots, v_{s_{q^-}}^-$.

Let $y=(y_1, y_2, \ldots, y_n)$ be an optimal solution of D_{F_0}. Then the following equations are satisfied:

$$\sum_{i=1}^{q} m_i u_r(x^i) + v_r^+ - v_r^- = \mu_r \qquad r=1, 2, \ldots, n. \tag{2.3}$$

$$\sum_{r=1}^{n} y_r u_r(x^i) = \phi(x^i) \qquad i=1, 2, \ldots, q. \tag{2.4}$$

$$y_{r_j} = F_1 \qquad j=1, 2, \ldots, q^+. \tag{2.5}$$

$$y_{s_k} = F_2 \qquad k=1, 2, \ldots, q^-. \tag{2.6}$$

REMARK: We have also: $q+q^++q^- \leq n$ and $m_i > 0$, $i=1, 2, \ldots, q$. The q column-vectors $u_1, u_2, \ldots, u_q$ in (2.3) given by $u_i=(u_1(x_i), u_2(x_i), \ldots, u_n(x_i))$ are linearly independent.

The relations (2.3), (2.4), (2.5), and (2.6) are necessary conditions for finding optimal solutions. They can be supplemented by further conditions.

Let $u_1, u_2, \ldots, u_n$ and ϕ have continuous partial derivatives of the first order. Define Q by

$$Q(x) = \sum_{r=1}^{n} y_r u_r(x) \qquad x \in S_k.$$

Then relation (2.4) takes the form

$$Q(x^i) = \phi(x^i) \qquad i=1, 2, \ldots, q.$$

If x belongs to S_k we have

$$Q(x) \geq \phi(x).$$

Let x^i be such that there is a nonzero vector h meeting the conditions:

$$x^i + h \in S_k, \qquad x^i - h \in S_k. \tag{2.7}$$

Define ψ by

$$\psi(t) = Q(x^i + th) - \phi(x^i + th) \qquad -1 \leq t \leq 1.$$

ψ has a continuous derivative with respect to t on $[-1, 1]$, $\psi(t) \geq 0$ and $\psi(0) = 0$. Hence $\psi'(0) = 0$.

This observation can be utilized to derive further constraints on Q in the following manner. Determine for each point x^i a system of linearly independent vectors h which meet the requirements (2.7). Denote these vectors by $h_1, h_2, \ldots, h_{l_i}$ (if there is none put $l_i = 0$). We have always $l_i \leq k$, the dimensionality of S_k. We note in passing that $l_i < k$ at boundary points only. Denote the directional derivative along h_j by D_j. Then we must conclude

$$D_j(Q(x_i) - \phi(x_i)) = 0 \qquad j = 1, 2, \ldots, l_i \qquad \text{if } l_j \geq 1. \tag{2.8}$$

$$i = 1, 2, \ldots, q$$

Equations (2.3), (2.4), (2.5), (2.6), and (2.8) form a nonlinear system and the optimal solution of the original problem can be found by solving it. The unknowns are the q masses $m_1, m_2, \ldots, m_q$ the n scalars $y_1, y_2, \ldots, y_n$, the vectors $x^1, x^2, \ldots, x^q$, and the numbers $v_1^+, v_2^+, \ldots, v_n^+, v_1^-, v_2^- \ldots, v_n^-$. We mention now two special cases: In the problems treated by Gustafson [20] and Gustafson-Rom-Kortanek [26], we have $k = 1$. Hence $l_j = 0$ at boundary points, $l_j = 1$ at interior points. If S_k is strictly convex (e.g., a nondegenerate ellipsoid), then $l_j = 0$ at boundary points and $l_j = k$ in the interior.

We next extend our results to nonlinear functions G. From Kortanek-Evans [34, p. 889], we get (after appropriate changes of notations):

LEMMA 2.6: Let G be a continuously differentiable function defined on an open convex set W in R^n. Consider the two problems:

(I)	(I_*)
$\min G(y)$	$\min y^T(\nabla G)_{y=y_*}$
when $y \epsilon K$	when $y \epsilon K$

where K is a closed convex set in W. Then y^* is optimal for I if and only if y^* is optimal for I^* provided either one of the following conditions hold

(a) G is pseudo-convex

(b) G is quasi-convex and $(\nabla G)_{y=y^*} \neq 0$.

Using this lemma, we realize that if G meets conditions (a) and (b) of the lemma, we can replace D_F by a linear problem with the objective function G^* defined by

$$G_*(y_1, y_2, \ldots, y_n) = \sum_{r=1}^{n} \frac{\partial G}{\partial y_r^*} y_r,$$

where $y_1^*, y_2^*, \ldots, y_n^*$ is an optimal solution of D_F. In our nonlinear system (2.3), (2.4), (2.5), (2.6),

(2.8), we should replace μ_r by $\frac{\partial G}{\partial y_r^*}$ in (2.3) in order to allow for nonlinear objective functions. (The remaining equations were derived independently of the objective function G.)

DEFINITION 2.3: We denote by *system NL* the nonlinear system of equations obtained by combining Equations (2.3) through (2.6) with (2.8). If G is nonlinear, μ_r in (2.3) should be replaced by $\frac{\partial G}{\partial y_r^*}$ as described above.

EXAMPLE 2.4:

Compute $$\min 4y_1 + 2/3\,(y_4 + y_6),$$

when $$y_1 + x_1 y_2 + x_2 y_3 + x_1^2 y_4 + x_1 x_2 y_5 + x_2^2 y_6 \geq 3 + (x_1 - x_2)^2 (x_1 + x_2)^2$$

$$S_2 = \left\{ (x_1, x_2) \;\middle|\; |x_i| \leq 1, \qquad i = 1, 2 \right\}.$$

The associated moment problem P_{F_0} reads

compute $$\max_{\alpha} \int_{S_2} [3 - (x_1 - x_2)^2 (x_1 + x_2)^2]\, d\alpha(x_1, x_2),$$

when $$\int_{S_2} d\alpha(x_1, x_2) = 4$$

$$\int_{S_2} x_1 d\alpha(x_1, x_2) = 0$$

$$\int_{S_2} x_2 d\alpha(x_1, x_2) = 0$$

$$\int_{S_2} x_1^2 d\alpha(x_1, x_2) = 2/3$$

$$\int_{S_2} x_1 x_2 d\alpha(x_1, x_2) = 0$$

$$\int_{S_2} x_2^2 d\alpha(x_1, x_2) = 2/3.$$

By inspection we find that P_{F_0} has the feasible solution with four mass points

m	x_1	x_2
1	$\frac{1}{\sqrt{6}}$	$\frac{1}{\sqrt{6}}$
1	$\frac{1}{\sqrt{6}}$	$-\frac{1}{\sqrt{6}}$

$$\begin{matrix} 1 & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 1 & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{matrix}$$

D_F has the feasible solution $y_1=3, y_2=y_3=\ \ldots\ =y_6=0$. The preference function assumes the value 12 in both problems, that is, we have found an optimal solution.

We observe that $d\alpha(x_1, x_2) = d(x_1, x_2)$ is another feasible solution of P_F and hence we have found the quadrature rule with positive weights:

$$\int_S \phi(x_1, x_2)d(x_1, x_2) = \phi\left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right) + \phi\left(\frac{-1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right) + \phi\left(\frac{1}{\sqrt{6}}, \frac{-1}{\sqrt{6}}\right) + \phi\left(\frac{-1}{\sqrt{6}}, \frac{-1}{\sqrt{6}}\right).$$

The rule has positive weights and is exact if ϕ is a polynomial of two variables and of degree less than 3.

3. COMPUTATIONAL SOLUTION

3.1. Definition of Acceptable Approximations

In this section we give the general principles of a computational scheme for the solution of the problem D_F (Definition 2.2).

DEFINITION 3.1: Let y be any vector. We define the *nonoptimality* $\Delta Z\ (y)$ as $|Z-G(y)|$, where Z is the optimal value of D_F.

DEFINITION 3.2: Let again y be any vector. We define the discrepancy of y by

$$\delta(y) = \min_x \left\{ \sum_{r=1}^{n} y_r u_r(x) - \phi(x) \right\} x \epsilon S_k.$$

Thus y is an optimal solution vector if $\delta(y) \geq 0$ and $G(y) = Z$. Generally, one has to be content with trying to find a vector y such that $\delta(y) \geq -\delta_0$ and $|G(y) - Z| \leq \epsilon_0$, where the positive numbers δ_0 and ϵ_0 are given tolerances. (Such a y is called an *acceptable approximation.*) We want to show that this can be done by means of a finite number of operations provided these are carried out with sufficiently good accuracy.

3.2 Cutting-plane Methods and Alternating Procedures

D_F amounts to minimize a convex function over a compact convex set K. This is not specified in the form of a few simple equations. Instead, we know the supporting planes which are:

$$\sum_{r=1}^{n} y_r u_r(x) - \phi(x) \geq 0 \qquad x \epsilon S.$$

One can, therefore contemplate using the principles of the cutting-plane method of Kelley [33] and Wolfe [51] with the accelerating device by Wolfe [51]. The first algorithm by Remez (see, e.g., Cheney [9], p. 96) can be considered as a variant of the cutting-plane method.

We generalize this algorithm and define the following *alternating procedure.* (The word alternating refers to the fact that the optimal solution of D_F is computed by alternatively minimizing G over subsets of K_F containing K and minimizing certain functions over S.

The general step is: let $x^1, x^2, \ldots, x^{s-1}$ $(s \geq 2)$ by given elements in S. Take y^s as an optimal solution vector of the problem

$$\min G(y),$$

subject to

$$\sum_{r=1}^{n} y_r u_r(x^j) - \phi(x^j) \geq 0 \qquad j = 1, 2, \ldots, s-1.$$

Then define x^s as an element in S which minimizes

$$\sum_{r=1}^{n} y_r^s u_r(x) - \phi(x) \qquad x \epsilon S.$$

If this last minimum is nonnegative, the process is terminated. Otherwise we generate $y^{s+1}, y^{s+2}, \ldots$.

THEOREM 3.1: The sequence $y^s, y^{s+1}, \ldots$ generated by the alternating procedure above converges toward an optimal solution of D_F.

PROOF: Since y^{s+1} meets all the constraints of y^s, $G(y^{s+1}) \geq G(y^s)$. The same is true for any optimal vector of D_F. Hence $G(y^s) \leq G(y^{s+1}) \leq \ldots \leq d$, where d is the optimal value. Three cases are conceivable, namely:

CASE A: The alternating process stops after a finite number of iterations.

CASE B: $\lim_{s \to \infty} G(y^s) = d - \eta, \qquad \eta > 0.$

CASE C: $\lim_{s \to \infty} G(y^s) = d.$

If Case A occurs, the optimal vector has been reached because the last vector satisfies the constraints of K. We want to show that Case B is not possible. Since y^j is an infinite sequence confined to the compact set K_F, it has accumulation points. Let y^* be such a point. Put $G(y^*) = d - \epsilon$, $\epsilon > 0$. Hence y^* does not belong to K.

Let $\bar{x}$ be an element in S which minimizes $\sum_{r=1}^{n} y_r^* u_r(x) - \phi(x)$. Denote the corresponding minimum by Δ. We must have $\Delta < 0$, since $y^* \notin K$.

From the definition of y^* we conclude

$$(*) \qquad \sum_{r=1}^{n} y_r^* u_r(x^j) - \phi(x^j) \geq 0 \qquad j = 1, 2, \ldots.$$

Let now $\{y^{lj}\}$ be a subsequence such that $y^{lj} \to y^*$ and x^{lj} tend toward an accumulation point x^*. We find for each j:

$$\sum_{r=1}^{n} y_r^{lj} u_r(\bar{x}) - \phi(\bar{x}) \geq \sum_{r=1}^{n} y_r^{lj} u_r(x^{lj}) - \phi(x^{lj}).$$

Letting $j \to \infty$ we arrive at

$$\Delta = \sum_{r=1}^{n} y_r^* u_r(\bar{x}) - \phi(\bar{x}) \geq \sum_{r=1}^{n} v_r^* u_r(x^*) - \phi(x^*).$$

But by (*)

$$\sum_{r=i}^{n} y_r^* u_r(x^*) - \phi(x^*) \geq 0 \text{ also,}$$

since $y_r^{lj} \to y_r^*$, $x^{lj} \to x^*$. This contradicts that $\Delta < 0$ and hence Case C is not possible. Theorem 3.1 is therefore proven.

The alternating method might be effective when an approximate solution vector of D_F is known.

If the objective function of D_F is linear, one can use the algorithms given in Gustafson [20] and Gustafson-Kortanek-Rom [26]. As a matter of fact, if the objective function is convex, D_F can be solved by solving a sequence of semi-infinite programs. We show now that the solution vector of D_F can be constructed as an accumulation point of the sequence $y^0, y^1, \ldots$ constructed recursively as follows.

Let y^0 belong to K. See def. 2.1.

When $y^0, y^1, \ldots, y^{l-1} (l=1, 2, \ldots)$ are determined, we define the linear functions

$$\pi_j(y) = G(y^j) + \sum_{r=1}^{n} (y - y_r^j) \left(\frac{\partial G}{\partial y_r}\right)_{y=y^j}$$

Then we define y^l as the optimal solution of the problem

$$\min \pi_{l-1}(y),$$

subject to

$$\pi_{l-1}(y) \geq \pi_j(y) \qquad j=0, 1, \ldots, l-2$$

$$\sum_{r=1}^{n} y_r u_r(x) \geq \phi(x) \qquad x \in S_K$$

$$\sum_{r=1}^{n} y_r u_r(x) \geq \phi(x) \qquad x \in S_F.$$

From Kelley [33] we conclude that $\{y^l\}$ contains a subsequence converging toward an optimal solution vector of D_F.

The methods discussed here can be of use to construct an approximate solution of the system NL, which is then solved by means of Newton-Raphson which is rapidly converging.

3.3 Approximation with Problems with a Finite Number of Constraints

3.3.1 Generalities

We first introduce:

DEFINITION 3.3: A finite subset T of S_K, $T = \{x^1, x^2, \ldots, x^N\}$ is called a *grid.*

DEFINITION 3.4: We denote by *problem* $D_F - T$ the task:

Compute $$\min G(y),$$

subject to

$$\sum_{r=1}^{n} y_r u_r(x^j) \geq \phi(x^j) \qquad x^j \epsilon T$$

$$\sum_{r=1}^{n} y_r u_r(x^j) \geq \phi(x) \qquad x \epsilon S_F.$$

(S_F is the same set as in the definition of problem D_F.) This problem can be solved by standard techniques of mathematical programming. The solution of $D_F - T$ can be used to approximate that of D_F. We shall now derive bounds for the discrepancy and nonoptimality expressed in grid data and characteristics of the functions $u_1, u_2, \ldots, u_n, \phi$ and G.

3.3.2 Error Bounds for Optimal Solutions of $D_F - T$

We need two definitions

DEFINITION 3.5: Let a grid T and a norm be given on S_k. The number $|T|$ given by

$$|T| = \max_{x \epsilon S_k} \min_{x_j \epsilon T} ||x - x_j||$$

will be called the *coarseness* of T.

This definition of $|T|$ agrees with the concept of "density" in Cheney [9, p. 84], but it is not consonant with the definition in Gustafson [20, p. 350]. The latter could not be directly extended to multidimensional grids. Since S_k is finite-dimensional, all norms are topologically equivalent and therefore any norm can be used.

DEFINITION 3.6: Continuity modulus of a real-valued function ψ. If ψ is continuous on S_k we define the function ω_ψ as follows:

$$\omega_\psi(z) = \sup |\psi(x') - \psi(x'')|,$$

subject to

$$x' \epsilon S_K, \qquad x'' \epsilon S_k$$

$$||x' - x''|| \leq z.$$

ω_ψ is called the *modulus of continuity* of ψ. ω_ψ is nonnegative and increasing. Since ψ is continuous $\lim \omega_\psi(z) = 0$ when $z \to +0$. (Compare Cheney [9, p. 86].)

We can now prove:

LEMMA 3.1: Let y^T be an optimal solution of $D_F - T$. Then

(3.1) $$\delta(y^T) \geq -\Delta_T(|T|), \text{ where}$$

(3.2) $$\Delta_T(|T|) = \sum_{r=1}^{n} |y_r^T| \omega_{u_r}(|T|) + \omega_\phi(|T|).$$

PROOF: (The arguments are a slight generalization of those in Gustafson [20, pp. 351–352].) Put

$$\psi = \sum_{r=1}^{n} y_r^T u_r - \phi.$$

Let $x \epsilon S_k$. We want to get a lower bound for $\psi(x)$. By the definition of $|T|$ there is a gridpoint x^j such that $||x - x^j|| \leq |T|$. We write

$$\psi(x) = \psi(x^j) + \psi(x) - \psi(x^j),$$

giving

$$\psi(x) \geq \psi(x^j) - |\psi(x) - \psi(x^j)|.$$

Hence

$$\psi(x) \geq \psi(x^j) - \omega_\psi(||x - x^j||) \geq -\omega_\psi(|T|)$$

since $\psi(x^j) \geq 0$, $||x - x^j|| \leq |T|$ and ω_ψ is positive and increasing. We find immediately

$$\omega_\psi(|T|) \leq \Delta_T(|T|). \qquad \text{Q.E.D.}$$

This result can be strengthened to

LEMMA 3.2: Let y^T be an optimal solution of $D_F - T$. Then

$$\delta(y^T) \geq -\Delta(|T|), \text{ where} \tag{3.3}$$

$$\Delta(|T|) = \omega_\phi(|T|) + F \sum_{r=1}^{n} \omega_{u_r}(|T|), \text{ and } F = \max\{|F_1|, |F_2|\}. \tag{3.4}$$

PROOF: Use the fact that $|y_r| \leq F$.

The bound in Lemma 3.2 is more conservative than Lemma 3.1, but it is *a priori* in the sense that we do not need to know y^T in order to evaluate it. Hence we can tell in advance how small $|T|$ must be selected in order to get a discrepancy below a given tolerance. We next derive bounds for nonoptimality.

LEMMA 3.3: Let y^T be an optimal solution of $D_F - T$. Let

$$H = \sum_{r=1}^{n} c_r u_r \tag{3.5}$$

be positive* over S_k and put

$$\gamma = \min_{x \epsilon S_k} H(x). \tag{3.6}$$

Put

$$y_r^H = y_r^T + \gamma^{-1} c_r \Delta_T(|T|), \qquad r = 1, 2, \ldots, n, \tag{3.7}$$

where Δ_T is defined by (3.2). Then

$$\left| \frac{1}{2}(G(y^T) + G(y^H)) - G(y) \right| \leq \frac{1}{2}(G(y^H) - G(y^T)). \tag{3.8}$$

where y is an optimal solution of D_F.

*The existence of H is guaranteed by Krein's condition.

PROOF: Put

$$Q^H = \sum_{r=1}^{n} y_r^H u_r.$$

We want to show that $Q^H(x) \geq \phi(x)$, $x \epsilon S_k$. We write

$$Q^H(x) - \phi(x) = Q^H(x) - Q^T(x) + Q^T(x) - \phi(x),$$

giving

$$Q^H(x) - \phi(x) = \gamma^{-1} H(x) \Delta_T(|T|) + Q^T(x) - \phi(x).$$

Thus

$$Q^H(x) - \phi(x) \geq \Delta_T(|T|)(\gamma^{-1} H(x) - 1) \geq 0.$$

Hence y^H is a feasible solution of D_F and therefore

$$G(y^H) \geq G(y).$$

On the other hand, y^T is the optimal solution of $D_F - T$ and we can therefore conclude

$$G(y^H) \geq G(y) \geq G(y^T),$$

from which (3.8) follows. Q.E.D.

Lemma 3.3 can be used to derive *a posteriori* bounds for nonoptimality. We can namely prove Lemma 3.4.

LEMMA 3.4: Let y^T be an optimal solution of $D_F - T$. Then we can replace Δ_T in (3.7) by Δ in (3.4) and (3.8) gives an *a priori* bound. Further, if v is such that

$$\left| \frac{\partial G}{\partial y_r} \right| \leq v_r \qquad r = 1, 2, \ldots, n$$

everywhere, then

$$\left| \frac{1}{2} (G(y^T) + G(y^H)) - G(y) \right| \leq \frac{1}{2\gamma} \Delta(|T|) \sum_{r=1}^{n} |c_r v_r|.$$

Often bounds of the partial derivatives of $u_1, \ldots, u_n$ and ϕ are known. Then the expressions for Δ and Δ_T can be simplified. Let x' and $x' + h$ belong to S. Then the mean-value theorem gives

$$|\phi(x' + h) - \phi(x')| = \sum_{r=1}^{n} \left(\frac{\partial \phi}{\partial x_r} h_r \right)_{x=\xi},$$

where $\xi = x' + \theta h$ for some number θ in (0, 1). Put

(3.9)
$$\kappa_\phi = \sup_{x \epsilon S} \sup_{h} \sum_{r=1}^{n} \left| h_r \frac{\partial \phi}{\partial x_r} \right|.$$

Hence

$$|\phi(x'+h)-\phi(x')| \leq ||h||\kappa_\phi, \text{ giving}$$

$$\omega_\phi(h) \leq ||h||\kappa_\phi.$$

Defining κ_{u_r} in the same manner as κ_ϕ, we obtain

$$\Delta_T(t) \leq t\Delta'_T \qquad t \geq 0, \tag{3.10}$$

$$\Delta(t) \leq t\Delta' \qquad t \geq 0 \tag{3.11}$$

where

$$\Delta'_T = \kappa_\phi + \sum_{r=1}^{n} |y_r|\kappa_{u_r}$$

$$\Delta' = \kappa_\phi + F\sum_{r=1}^{n} \kappa_{u_r}, \text{ and } F = \max\{|F_1|, |F_2|\}.$$

If we replace Δ_T and Δ with the bounds (3.10) and (3.11) and revise the arguments in the preceding four lemmas, we arrive at Theorem 3.2.

THEOREM 3.2: If $u_1, u_2, \ldots, u_n$ and ϕ have bounded partial derivatives of the first order and y^T is an optimal solution of $D_F - T$, one can give explicit *a priori* bounds on the nonoptimality and discrepancy of y^T. These bounds are proportional to $|T|$.

A further refinement is possible if $u_1, u_2, \ldots, u_n$ and ϕ have continuous partial derivatives of the second order.

Let y^T be an optimal solution of $D_F - T$. Put

$$\psi = \sum_{r=1}^{n} y_r u_r - \phi.$$

Then $\psi(x^j) \geq 0$, for all $x^j \epsilon T$. A lower bound of $\psi(x)$, $x \epsilon S_k$ can be constructed from the following result.

LEMMA 3.5: Let $x^1, x^2, \ldots, x^{k+1}$ be $k+1$ given points in S_k and h a number such that

$$0 \leq |x_i^r - x_i^r| \leq h \qquad i = 1, 2, \ldots, k; \qquad r = 1, 2, \ldots, k+1$$

and the determinant

$$\begin{vmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_k^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_k^2 \\ 1 & x_1^3 & x_2^3 & \cdots & x_k^3 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & x_1^{k+1} & x_2^{k+1} & & x_k^{k+1} \end{vmatrix} \neq 0.$$

Take a fixed point x in the convex hull U of $x^1, x^2, \ldots, x^{k+1}$. Put $R(x) = \sup f(x)$, when f varies over all functions with two continuous partial derivatives of the second order such that

$$\left| \frac{\partial^2 f}{\partial x_i \partial x_j} \right| \leq 2c_{ij} \qquad x \epsilon U, \tag{3.12}$$

where c_{ij} $i=1, 2, \ldots, k$; $j=1, 2, \ldots, k$ are given constants and f meets the condition

$$f(x^j) = 0 \qquad j=1, 2, \ldots, k+1, \tag{3.13}$$

then

$$R(x) \leq \sum_{i=1}^{k} \sum_{j=1}^{k} c_{ij} \left(|x_i x_j| + \sum_{r=2}^{k+1} \lambda_r |x_i^r x_j^r| \right) \tag{3.14}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_{k+1}$ are determined by

$$x = \sum_{r=1}^{k+1} \lambda_r x^r.$$

REMARK. To determine $R(x)$ over a fixed x is an instance of problem D in section 1 when S is infinite-dimensional. S is namely the space of all functions of k variables which have continuous partial derivatives of the second order.

PROOF: Without loss of generality, we can assume that the coordinates are chosen such that $x^1 = 0$. Taylor's formula (expanding about x^1) gives, since $f(x^1) = 0$:

$$f(x^r) = \sum_{j=1}^{k} a_j x_j + \sum_{i=1}^{k} \sum_{j=1}^{k} b_{ij} x_i x_j, \qquad r=2, 3, \ldots, k+1.$$

where

$$a_j = \left(\frac{\partial f}{\partial x_j} \right)_{x=x^1}$$

$$b_{ij}^r = \frac{1}{2} \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{x=\theta_r x^1 + (1-\theta_r) x^r} 0 \leq \theta_r \leq 1.$$

Hence we arrive at the problem:

Compute $$L = \sup \sum_{j=1}^{k} a_j x_j + \sum_{i=1}^{k} \sum_{j=1}^{k} b_{ij} x_i x_j,$$

subject to

$$\sum_{j=1}^{k} a_j x_j^r + \sum_{i=1}^{k} \sum_{j=1}^{k} b_{ij}^r x_i^r x_j^r = 0 \qquad r = 2, 3, \ldots, k+1 \tag{3.15}$$

$$|b_{ij}| \leq c_{ij}, \; |b_{ij}^r| \leq c_{ij}.$$

(3.15) gives $$\sum_{j=1}^{k} a_j x_j^r = -\sum_{i=1}^{k}\sum_{j=1}^{k} b_{ij}^r x_i^r x_j^r.$$

Since x is in the convex hull of $\{x^r\}$ $r = 1, 2, \ldots, k+1$, we can write

$$x = \sum_{r=1}^{k+1} \lambda_r x^r \text{ where } \sum_{r=1}^{k+1} \lambda_r = 1, \qquad \lambda_r \geq 0, \qquad r = 1, 2, \ldots, k+1,$$

giving

$$\sum_{j=1}^{k} a_j x_j = -\sum_{r=2}^{k+1} \lambda_r \sum_{i=1}^{k}\sum_{j=1}^{k} b_{ij}^r x_i^r x_j^r.$$

Thus we are left with the task

Compute $$L = \sup \sum_{i=1}^{k}\sum_{j=1}^{k} \left(b_{ij}\, x_i x_j - \sum_{r=1}^{k+1} \lambda_r x_i^r x_j^r \right),$$

subject to $$|b_{ij}^r| \leq c_{ij}.$$

Hence we should take

$$b_{ij} = c_{ij} \operatorname{sign}\left(x_i x_j - \sum_{r=1}^{k+1} \lambda_r x_i^r x_j^r \right)$$

Entering $|b_{ij}| = |b_{ij}^r| = c_{ij}$, we get the bound sought. The determinant condition implies that $\lambda_1, \lambda_2, \ldots, \lambda_{k+1}$ are uniquely determined by x and $x^1, x^2, \ldots, x^{k+1}$, and hence the lemma is proven.

If we now make the substitutions $x^r = h\xi^r$, $x = h\xi$ in (3.14), we get

$$R(x) \leq h^2 \sum_{i=1}^{k}\sum_{j=1}^{k} \left(c_{ij} \left| \xi_i \xi_j \right| + \sum_{r=1}^{k+1} \left| \lambda_r \xi_i^r \xi_j^r \right| \right),$$

where

$$\xi = \sum_{r=1}^{k+1} \lambda_r \xi^r.$$

If now the bounds on the derivatives hold uniformly over S_k, we get

$$\delta(y^T) \geq -e \cdot |T|^2,$$

where e is determined by the distribution of the grid points T and the bounds of the second-order partial derivatives on S_k. Hence, if we consider a sequence $T_1, T_2, \ldots$ of hypercubic grids the bound on $|\delta(y^{T_i})|$ decreases as the square of $|T_i|$, $i = 1, 2, \ldots$. Revising the arguments leading to Theorem 3.2 above, we find that the same holds true for the bound on the nonoptimality.

3.3.3 Convergence Results when $|T| \to 0$

LEMMA 3.6: To every $\epsilon > 0$ there is an $h > 0$ such that if $|T| < h$ and y^T is an optimal solution of $D_F - T$, there is a y which is an optimal solution of D_F and satisfies $|| y - y^T || < \epsilon$.

PROOF: The same arguments apply as in Gustafson [20, Theorem 3.3]. This result can be both generalized and sharpened.

If $D_F - T$ has an optimal solution y^T, we can associate with it the problem $P_F - T$ given below.

DEFINITION 3.7: Let the grid T be $\{x^1, x^2, \ldots, x^N\}$. We denote by problem $P_F - T$ the task:

Compute
$$\max \sum_{j=1}^{N} m_j \phi(x^j) + \sum_{r=1}^{n} (F_1 v_r^+ - F_2 v_r^-),$$

subject to
$$\sum_{j=1}^{N} m_j u_r(x^j) + v_r^+ - v_r^- = \left(\frac{\partial G}{\partial y_r}\right)_{y^T}$$

$$m_j \geq 0 \qquad v_r^+ \geq 0 \qquad v_r^- \geq 0.$$

$P_F - T$ is a linear program (even if G is not linear) and has hence an optimal solution which corresponds to a point mass distribution with, at most, n mass-points. Select such a solution of $P_F - T$ with the minimum number of mass-points. Then with each T we can associate the vector $z(T)$ given by

(3.16)
$$z(T) = \xi^1(T), \xi^2(T), \ldots, \xi^v(T), y(T),$$

where ξ^j are the mass-carrying points. An optimal solution of $P_F - T$ and $D_F - T$ is uniquely determined if $z(T)$ is given, since the vectors $\xi^1(T), \ldots, \xi^v(T)$ are linearly independent and v_r^+, v_r^- enter the solution if $y_r = F_1$ or F_2, respectively.

Let $||\ \ ||$ be a norm on S_k. We define $||z||$ by

$$||z|| = \sum_{j=1}^{v} ||\xi^j|| + \sum_{r=1}^{n} |y_r|.$$

THEOREM 3.3: Let T_j $j = 1, 2, \ldots$ be a sequence of grids such that $|T_j| \to 0$ when $j \to \infty$. With each T_j we associate the vector $z(T_j)$ defined analogously with $z(T)$ above. Then we can find a subsequence $z(T_{j_1})$ converging towards a vector z which describes a solution of $P_F - D_F$.

PROOF: Since $v \leq n$, $z(T)$ has at most $n(k+1)$ components. S_k is a compact set and $F_1 \leq y_r \leq F_2$. Hence we can find a number B such that $||z(T_j)|| \leq B$. Therefore $\{z(T_j)\}$ is confined to a bounded subset of a finite-dimensional Banach space. Thus $\{z(T_j)\}$ is contained in a compact set. We first select a subsequence $T_{j_1}, T_{j_2}, T_{j_n}$ such that $y(T_{j_s})$ converges towards a vector $\bar{y}$. Using the same arguments as in Theorem 3.3 in Gustafson [20], we establish that $\bar{y}$ is an optimal solution of D_F.

To each $y(T_{j_s})$ there is an optimal solution of $P - T_{j_s}$. We want to define vectors $z(T_{j_s})$ according to (3.16) which we do recursively as follows:

Let $z(T_{j_s})$ be given and of the form

$$z(T_{j_s}) = \xi^1(T_{j_s}), \xi^2(T_{j_s}), \ldots, \xi^v(T_{j_s}), y(T_{j_s}).$$

Let $P_F - T_{j_{s+1}}$ have an optimal solution with the mass-carrying points $u^1, u^2, \ldots, u^{\bar{v}}$. We now define $\xi^l(T_{j_{s+1}})$ equal to a vector from the set $\{u^1, u^2, \ldots, u^{\bar{v}}\}$ which minimizes

$$||\xi^l(T_{j_s}) - u^\sigma||,$$

when $1 \leqslant \sigma \leqslant \bar{v}$.
This is done for $l = 1, 2, \ldots, \min(v, \bar{v})$.
There are three cases

(a) $\bar{v} < v$. Then the vectors in $z(T_{j_s})$ which have not been matched are put in the vector for $z(T_{j_{s+1}})$.

(b) $\bar{v} = v$. No subsequent change in the definition of $z(T_{j_{s+1}})$ is done.

(c) $\bar{v} > v$. The vectors from $\{u^1, u^2, \ldots, v^{\bar{v}}\}$ which have not been matched are transferred to $z(T_{j_{s+1}})$. Hence, in all cases max $(v, \bar{v})$ vectors are put in the vector for $z(T_{j_{s+1}})$ which also contains $v(T_{j_{s+1}})$.

In the manner described above, we define recursively $z(T_{j_1}), z(T_{j_2}), \ldots$. In no case will any of these vectors contain more than n points.

We can now take a subsequence which converges towards an accumulation point which we call z^* from which a solution of $D_F - T$ and its corresponding point-masses can be constructed.

REMARK: From the construction of z^* it is obvious that certain of the points represented in the vector z^* carry the mass zero. They can hence be removed. Other points can be confluent. If this is the case, only one member from every group of confluent points is carried. The last mentioned case is common. Compare Gustafson [20] and subsection 3.3.5.

The idea to approximate a semi-infinite program with an optimization problem with a finite number of constraints is, of course, not new. Convergence results of the same character can be found, e.g., in Vershik-Temel't [48] and Cheney [9, pp. 86–88].

3.3.4 Special Devices to Economize and Stabilize the Simplex Method

In this subsection we consider the case when G is linear, that is

$$G(y) = \sum_{r=1}^{n} y_r \mu_r.$$

(As noted in section 3.2, every convex semi-infinite program can be solved by solving a sequence of linear semi-infinite programs.)

In this case, $D_F - T$ is a linear program and an optimal vector y^T can be constructed with the Simplex method.

Let now y be a candidate for y^T and put

$$\psi_y = \sum_{r=1}^{n} y_r u_r - \phi. \tag{3.17}$$

We want to investigate if $\psi_y(x^j) \geqslant 0$, $x^j \epsilon T$. Let $x^i \epsilon T$ be such that $\psi_y(x^i) > 0$. For all x in S_k we find the bound

$$\psi_y(x) \geqslant \psi_y(x^i) - \omega_{\psi_y}(||x - x^i||).$$

We note that

$$\omega_{\psi_y}(t) \leq \sum_{r=1}^{n} |y_r| \omega_{u_r}(t) + \omega_{\phi(y)}, \qquad t > 0.$$

If we put

$$S_i = \{x | \omega_{\psi_y}(||x - x^i||) \leq \psi(x^i)\},$$

we can conclude that $x \epsilon S_i$ implies $\psi_y(x) \geq 0$. Thus in particular, if $x^j \epsilon T$ is in S_i the corresponding column need not be even generated in order to establish that $\psi(x^j) \geq 0$.

Another problem is that numerical difficulties can be anticipated when $|T|$ is small, due to the fact that the Simplex algorithm calls for the solution of a linear system, whose matrix of coefficients might be nearly singular. Each Simplex iteration consists of two stages:

A. Determination of a candidate vector y by solution of the system

$$\sum_{r=1}^{n} y_r u_r(x^j) = \phi(x^j) \qquad j = 1, 2, \ldots, n, \tag{3.18}$$

where x^j belongs to T and corresponding rows are linearly independent. Then we have to find the sign of $\psi_y(x^i)$ for all $x^i \epsilon T$, where ψ_y is defined by (3.17).

B. If $\min \psi(x^i)$, $x^i \epsilon T$, is assumed for $i = i_1$, we introduce x^{i_1} in the next basis and hence we have to solve systems of the form

$$\sum_{i=1}^{n} m_i u_r(x^i) = \mu_r \qquad r = 1, 2, \ldots, n \tag{3.19}$$

$$m_i > 0.$$

As remarked in Gustafson [20], the abscissae often lie close together in pairs which will cause the matrix of coefficients in (3.18) and (3.19) to be ill-conditioned.

Consider first the problem of determining $\psi_y(x)$ from (3.18). Let $\bar{y}$ be the computed value of y and put

$$\epsilon_j = \left| \sum_{r=1}^{n} \bar{y}_r u_r(x^j) - \phi(x^j) \right| \qquad j = 1, 2, \ldots, n.$$

Let further $\Delta\psi_y(x)$ be the error in the value of $\psi_y(x)$ caused by the fact that we use $\bar{y}$ instead of y.

Gustafson [22] gives the bound

$$|\Delta\psi_y(x)| \leq \sum_{j=1}^{n} \epsilon_j |\rho_j(x)|,$$

where

$$\sum_{j=1}^{n} \rho_j(x) u_r(x^j) = u_r(x) \qquad r = 1, 2, \ldots, n.$$

Put

$$\bar{\rho}(x) = \sum_{r=1}^{n} |\rho_r(x)|.$$

We note then that $\bar{\rho}(x)$ is a continuous function over S_k and that $\bar{\rho}(x_j)=1, j=1, 2, \ldots, n$. Further, $\bar{\rho}$ is completely determined by the system $u_1, u_2, \ldots, u_n$ and hence independent on the way we perform our computations.

ϵ_r is, however, dependent both on the computer and the manner in which (3.18) is solved. Let $\bar{\epsilon}_r$ be the value we should obtain if we inserted the exact vector y in (3.18) and evaluated the residuals computationally. Wilkinson [50, p. 252] states that if (3.18) is solved by means of Gaussian elimination with pivoting, then $||\epsilon|| \leq 3||\bar{\epsilon}||$ even if the system is ill-conditioned. Therefore, if we solve (3.18) in this mode, $|\Delta\psi_y(x)|$ is as small as possible. However, the ordinary Simplex method does not provide pivoting when the sequence of linear systems is solved, a fact that may be the cause of the often-observed instability of linear programming codes. In contrast, the variant of Bartels-Golub-Saunders [52] holds promise to be more stable.

When we solve (3.19), it is crucial that the sign of m_i remains positive.

3.3.5. Construction of an Initial Approximation for Newton-Raphson Method

We discuss now how to construct an initial approximation for system *NL* (Definition 2.3), when the solution of D_F-T and P_F-T is known. In this subsection we make the following *general assumptions:*

A1: $u_1, u_2, \ldots, u_n$ and ϕ have continuous partial derivatives of the second order.

A2: G is linear

$$G(y)=\sum_{r=1}^{n}\mu_r y_r. \tag{3.20}$$

(If G has continuous partial derivatives, we put

$$G(y)\approx G(y^*)+\sum_{r=1}^{n}(y_r-y_r^*)\left(\frac{\partial G}{\partial y_r}\right)_{y_r^*}, \tag{3.21}$$

where y_r^* is a solution of D_F-T.)

A3: The matrix $A_y(x)$ given by

$$A_y^{ij}(x)=\frac{\partial^2\psi_y(x)}{\partial x_i\partial x_j} \tag{3.22}$$

is positive definite when x is a zero of ψ_y, and ψ_y is defined by (3.17).

REMARK: The linearization (3.21) is used when we employ the iterative process described in the end of subsection 3.2. Assumption 3 entails that the zeroes correspond to strict minima of ψ_y. A3 is difficult to verify in advance.

The major problem in finding an approximate solution is to determine the number of mass-points in an optimal solution of D_F when the solution of D_F-T and its primal P_F-T are known. Let y^T be an optimal solution of D_F-T and an optimal solution of P_F-T be described by the pairs

$$(\xi^i, m^i), \qquad i=1, 2, \ldots, n'. \tag{3.23}$$

*Research Report, Department of Computer Sciences, Standford, 1969.

The vector ξ^i gives the location of the mass m^i. $P_F - T$ may have many optimal solutions, but we can always take one with $n' \leq n$.

DEFINITION 3.8: A subset C_j of $\{\xi^i\}_{i=1}^{n}$ is called a *cluster* if each member of C_j lies at most $3|T|$ from any other member of C_j, and C_j cannot be expanded by inclusion of more elements of $\{\xi^i\}_{i=1}^{n'}$.

Thus we divide the set (3.23) uniquely in q clusters where $1 \leq q \leq n'$.

DEFINITION 3.9: Put

$$\psi_T = \sum_{r=1}^{n} y_r^T u_r - \phi$$

and define the matrix A_T analogously with A_y in (3.22).

C_i is called a *point-group* if $A_T(x^j)$ is positive definite and

$$\|\nabla \psi_T(x^j)\| \leq 0.5 \|A_T(x^j)\|, \qquad \text{all } x^j \epsilon C^i.$$

Two cases are possible, namely: case a: all clusters are point-groups; case b: there is a cluster that is not a point-group.

LEMMA 3.7: Let $A1, A2, A3$ hold. Then there is a number h' such that if $|T| < h'$ then all clusters are point-groups.

PROOF: Assume the contrary. Using Theorem 3.3, we can then select a sequence $T_1, T_2, \ldots$ of grids such that $|T_l| \to 0$ and such that corresponding vectors $z(T_l)$ tend to a vector z while at least one of the clusters of $z(T_l)$ is not a point-group. Let the clusters be $C_1(l), C_2(l), \ldots, C_{q(l)}(l)$. Each of the clusters contains at most n elements and hence their diameter is less than $3n|T_l|$. Hence all the mass-carrying points in a cluster converge toward the same point. Denote these limit-points by $\bar{\xi}_j$, $j = 1, 2, \ldots, q$. Using Assumption $A3$, we conclude that there is a $\delta_0 > 0$ such that if $\|x - \bar{\xi}_j\| < \delta_0$, $j = 1, 2, \ldots, q$ then $\|A(x)\| \geq 4\|\nabla \psi(x)\|$. The convergence of $y_1^1, y_2^2, \ldots$, implies that there is an N_1 such that $\|A_{T_l}(x)\| \geq 2\|\nabla \psi_{T_l}\|$, $l > N_1$ for $\|x - \bar{\xi}\| < \delta_0$. In the same manner we establish that there is an N_2 such that $l > N_2$ implies $A_{T_l}(x)$ is positive definite. Therefore, if $l > \max(N, N_1, N_2)$ then C_j is a point-group, $j = 1, 2, \ldots, q$. Hence the sought contradiction is established.

If T is such that the set (3.23) can be subdivided in clusters, all of which are point-groups, we construct an initial approximation to system NL as follows:

1. The masses in each point-group are combined and allocated in the group's center of gravity. Hence we take q equal to the number of point-groups and each point-group corresponds to a mass-point.
2. If a mass-point is less than $3|T|$ from the boundary, it is moved to the nearest boundary-point.
3. This so obtained point-mass distribution is taken as first approximation.
4. Equations (2.4) and (2.8) are determined by the distribution of mass-points.

3.3.6 Use of Newton-Raphson Methods

Applying the methods described in the preceding section, we obtain an approximate solution of the system NL.

A solution of this system can be described by a vector z of the general structure

$$z = (m_1, x^1, m_2, x^2, \ldots, m_q, x^q, v^+, v^-, y), \tag{3.24}$$

where the point x^i carries the mass m_i and y is an optimal solution of D_F.

DEFINITION 3.10: Any vector of the general structure (3.24) will be called a *trial* vector if $m_i > 0$, $i = 1, 2, \ldots, q$, $v_r^+ \geq 0$, $r = 1, 2, \ldots, n$ and $v_r^- \geq 0$, $r = 1, 2, \ldots, n$.

LEMMA 3.8: Any solution of (2.3)–(2.6), (2.8) which is a trial vector, can be used to give a lower bound of the optimal value of D_F. If also

$$\sum_{r=1}^{n} y_r u_r(x) \geq \phi(x), \qquad x \in S_k,$$

then y is an optimal solution of D_F.

PROOF: A trial vector which is a solution describes a feasible solution of $P_F - T$ for a certain T. The conclusions follow, then, from known duality relations. Q.E.D.

We want to generate a sequence of trial vectors which converge toward an optimal solution. We write the system NL under the general form

$$W(z) = 0.$$

Assume a trial vector z^j is known. Then we want to construct a correction h^j such that

$$\|W(z^j + h^j)\| < \|W(z^j)\|, \tag{3.25}$$

and then put

$$z^{j+1} = z^j + h^j, \tag{3.26}$$

if h^j can be selected such that $z^j + h^j$ is a trial vector.

In the classical Newton-Raphson method, we take h^j as the solution b^j of the system.

$$\sum_{r=1}^{N} \frac{\partial W(z^j)}{\partial z_r} b_r^j + W(z^j) = 0 \tag{3.27}$$

where N is the number of components of z. If the matrix in (3.27) is singular, the methods in Ben-Israel [2] can be used.

In Ortega-Rheinboldt [40, p. 421], we find general criteria for the convergence of the Newton-Raphson's method, but they cannot, in general, be used. However, if the matrix in (3.27) is regular for all z^j and the condition (3.25) is met, then $\{z_j\}$ converges toward a local minimum of the function $\|W(z)\|$. The same is true for the modified sequence $\{z^j\}$ obtained by putting $h^j = \lambda \cdot b^j$ where the real number λ is such that condition (3.25) is met together with the requirement that z^{j+1} is a trial vector.

The general idea is to generate a sequence of trial vectors until the norm of the residual falls below a value, prescribed in advance. If this does not happen before a given maximum time has elapsed, the process is assumed to diverge. Then one can use the last trial vector found to construct better approximations by means of the grid point methods described in earlier sections.

We note that D_F subsumes a large class of different problems. In many important particular instances, special methods can be used both to simplify the computational scheme and to establish properties of convergence and unicity. We will return to this in later papers.

3.3.7 Remarks on Sensivity

It is well known from the numerical solution of special cases of D_F (see, e.g., Example 1.3 in the introduction) that small changes in input data cause large dislocations of x^i and m_i, but that the optimal value is not affected very much. x^i are the locations of minima of the function

$$\sum_{r=1}^{n} y_r u_r - \phi,$$

and to determine these is an ill-conditioned task in one dimension (see, e.g., Wilkinson [49, p. 39]) and this situation cannot be expected to improve when S_k has several dimensions.

A first order *a posteriori* approximation of the sensivity can be made if the matrix of (3.27) is regular. Let $f(z)=G(z)$ and let $\bar{z}$ be an approximation of a solution z^*. Linearizing, we arrive at

$$df=f(z^*)-f(\bar{z}) \approx <\nabla f,\ b>,$$

where

$$Mb=-W(\bar{z})$$

with

$$M^{ij}=\frac{\partial W_i(\bar{z})}{\partial z_j} \text{ and } <a, b> \text{ denotes } \sum_{r=1}^{N} a_r b_r.$$

Hence we find $\|b\|=\|M^{-1}W(\bar{z})\| \leq \|M^{-1}\|\,\|W(\bar{z})\|$ and we get the approximation $\|z-z^*\| \approx \|b\|$. The estimate of df can be written in an attractive manner. Using Lemma 2.1 in Gustafson [22], we get $df \approx -<u,\ w>$ where $M^T u=\nabla f$. Hence an approximate bound on $|df|$ is given by $|<u, w>|$. This concludes our treatment of the computational scheme.

Recently (1973) a definition of cluster is taken independently of grid size. The convergence theorems are then different than lemma 3.7.

BIBLIOGRAPHY

[1] Baker, George A., Jr. and John L. Gammel, "Applications of the Principle of the Minimum Maximum Modulus to Generalized Moment Problems and Some Remarks on Quantum Field Theory," J. Math. Anal. and Appl. *33*, 197–211 (1971).

[2] Ben-Israel, A., "A Newton-Raphson Method for the Solution of Systems of Equations," J. Math. Anal. and Appl. *15*, 243–252 (1966).

[3] Ben-Israel, A., A. Charnes, and K. O. Kortanek, "Asymptotic Duality Over Closed Convex Sets," J. Math. Anal. and Appl. *35* (1971).

[4] Bojanic, R. and R. DeVore, "On Polynomials of Best One-Sided Approximation," L'Ensignement Math. *12*, 139–164 (1966).

[5] Buck, R. C., "Alternation Theorems for Functions of Several Variables," J. Approx. Theory *1*, 325–334 (1968).

[6] Charnes, A. and W. W. Cooper, *Management models and industrial applications of linear programming* (J. Wiley and Sons, New York: 1961) Vols. I and II.

[7] Charnes, A., W. W. Cooper, and K. O. Kortanek, "Duality, Haar Programs and Finite Sequence Spaces," Proc. Nat. Acad. Sci. U.S., *48*, 783–786 (1962).

[8] Charnes, A., W. W. Cooper, and K. O. Kortanek, "On the Theory of Semi-Infinite Programming and a Generalization of the Kuhn-Tucker Saddle Point Theorem for Arbitrary Convex Functions," *NRLQ 16*, 41–51 (1969).

[9] Cheney, W. E., *Introduction to approximation theory* (McGraw-Hill, Inc., N.Y., 1966).

[10] Dantzig, G. B., *Linear programming and extensions* (Princeton University Press, N.J., 1963).

[11] DeVore, R., "One Sided Approximations of Functions," J. Approx. Theory *1*, 11–25 (1968).

[12] Duffin, R. J., "Infinite Programs," in Linear Inequalities and Related Systems (ed. by H. W. Kuhn and A. W. Tucker), Annals of Math. Studies No. 38, Princeton University Press, Princeton, N.J., pp. 157–170 (1956).

[13] Duffin, R. J., "An Orthogonality Theorem of Dines Related to Moment Problems and Linear programming," J. Combinatorial Theory *2*, 1–26 (1967).

[14] Duffin, R. J., "Duality Inequalities of Mathematics and Science," 401–423 in [39].

[15] Duffin, R. J. and L. A. Karlovitz, "Formulation of Linear Programs in Analysis I: Approximation Theory," SIAM Jour. *16*, 662–675 (1968).

[16] Fan, Ky, "On Systems of Linear Inequalities," in *Linear Inequalities and Related Systems* (ed. by H. W. Kuhn and A. W. Tucker), Annals of Math. Studies No. 38, Princeton University Press, Princeton, N.J. (1956), pp. 99–156.

[17] Fan, Ky, "Asymptotic Cones and Duality of Linear Relations," J. Approx. Theory *2*, 152–159 (1969).

[18] Gorr, W., and K. O. Kortanek, "Numerical Aspects of Pollution Abatement Problems: Constrained Generalized Moment Techniques," IPP Report No. 12, School of Urban and Public Affairs, Carnegie-Mellon University (Oct. 1970).

[19] Gorr, W., S.-Å. Gustafson, and K. O. Kortanek, "Optimal Control Strategies for Air Quality Standards and Regulatory Policy," Environment and Planning *4*, 183–192, (1972).

[20] Gustafson, S.-Å., "On the Computational Solution of a Class of Generalized Moment Problems," SIAM J. Numer. Analysis *7*, 343–357 (1970).

[21] Gustafson, S.-Å., "Numerical Aspects of the Moment Problem," Fil.dr. Thesis, Institutionen för Informations Behandling, Stockholms Universitet, Stockholm, Sweden (Apr. 1970).

[22] Gustafson, S.-Å., "Control and Estimation of Computational Errors in the Evaluation on Interpolation Formulae and Quadrature Rules," Math. Computation *24*, 847–854 (1970).

[23] Gustafson, S.-Å. and Germund Dahlquist, "On the Computation of Slowly Convergent Fourier Integrals," Methoden und Verfahren der Mathematischen Physik *6*, 37–43 (1972).

[24] Gustafson, S.-Å. and K. O. Kortanek, "Analytical Properties of Some Multiple-Source Urban Diffusion Models," Environment and Planning, *4*, 31–41, (1972).

[25] Gustafson, S.-Å. and K. O. Kortanek, "Mathematical Models for Air Pollution Control: Numerical Determination of optimizing Abatement Policies" to appear in *Models for Environmental Pollution Control* (R. A. Deininger, Ed.), Ann Arbor Science Press, Ann Arbor, Mich.

[26] Gustafson, S.-Å., K. O. Kortanek, and W. Rom, "Non-Čebyševian Moment Problems," SIAM J. Numer. Analysis *7*, 335–342 (1970).

[27] Gustafson, S.-Å. and J. Martna, "Numerical Treatment of Size Frequency Distributions with Computer Machine," Geologiska Foreningens Forhandlingar *84*, 372–389 (1962).

[28] Gustafson, S.-Å. and W. Rom, "Applications of Semi-Infinite Programming to the Computa-

tional Solution of Approximation Problems," Tech. Report No. 88, Dept. of Operations Research, Cornell University, Ithaca, N.Y. (Sept. 1969).

[29] Haar, A., "Über lineare ungleichungen," Acta. Math. (Szeged) *2*, 1–14 (1924).

[30] John, Fritz, "Extremum Problems with Inequalities as Side Conditions, in: *Studies and essays*, Courant Anniversary Vol. (ed. K. O. Friedrichs, O. E. Neugebauer, and J. J. Stoker) J. Wiley and Sons, Inc., New York, pp. 187–204 (1948).

[31] Kantorovich, L. V. and G. Sh. Rubinshtein, "Concerning a Functional Space and Some Extremum Problems," Dokl. Akad. Nauk. *SSSR 115*, 1058–1061 (1957).

[32] Karlin, S. and W. J. Studden, *Tchebycheff Systems: with Applications in Analysis and Statistics* Interscience Publishers, J. Wiley and Sons, Inc., New York, (1966).

[33] Kelley, J. E., Jr., "The Cutting Plane Method for Solving Convex Programs," J. SIAM *8*, 703–712 (1960).

[34] Kortanek, K. O. and J. P. Evans, "Pseudo-Concave Programming and Lagrange Regularity," Operations Research *15*, 882–891 (1967).

[35] Krafft, Olaf, "Programming Methods in Statistics and Probability Theory," 425–446 in [39].

[36] Kretschmer, K. S., "Programmes in Paired Spaces," Can. J. Math. *13*, 221–238 (1961).

[37] Kretschmer, K. S., "Linear Programming in Locally Convex Spaces and Its Use in Analysis," Ph.D. Thesis, Carnegie-Mellon University, Pittsburgh, Pa. (1958).

[38] Meinardus, Günter, *Approximations of functions: theory and numerical methods* (Springer-Verlag, New York, Inc., 1967).

[39] *Nonlinear programming* (ed. J. B. Rosen, O. L. Mangasarian, and K. Ritter) (Academic Press, New York, 1970).

[40] Ortega, J. M. and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables* (Academic Press, New York and London, 1970).

[41] Powell, M. J. D., "On the Maximum Errors of Polynomial Approximation Defined by Interpolation and by Least Square Criteria," Comp J. *9*, 404–407 (1966).

[42] Rivlin, T. J. and H. S. Shapiro, "A unified Approach to Certain Problems of Approximation and Minimization," SIAM J. Appl. Math. *9*, 670–699 (1961).

[43] Rubinshtein, G. Sh., "Investigations on Dual Extremal Problems," Doctoral Dissertation, Inst. Matem. SO AN SSSR, Novosibirsk (1965).

[44] Shapiro, H. S., "On a Class of Extremal Problems for Polynomials in the Unit Circle," Portugaliae Math. *20*, 67–93 (1961).

[45] Shohat, J. A. and J. D. Tamarkin, "The Problem of Moments," Mathematical Surveys, No. 1, Am. Math. Soc., New York (1943).

[46] Stiefel, E., "Note on Jordan Elimination, Linear Programming and Tchebycheff Approximation," Numer. Math. *2*, 1–17 (1960).

[47] Todd, J., *A survey of numerical analysis* (McGraw-Hill, New York, 1962).

[48] Vershik, A. M. and V. Temel't, "Some Questions Concerning the Approximation of the Optimal Value of Infinite-Dimensional Problems in Linear Programming," Sibirskii Matematicheskii Zhurnal *9*, 591–601 (1968).

[49] Wilkinson, J. H., *Rounding Errors in Algebraic Processes* (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963).

[50] Wilkinson, J. H., *The Algebraic Eigenvalue Problem* (Clarendon Press, Oxford, 1965).

[51] Wolfe, Philip, "Accelerating the Cutting Plane Method for Nonlinear Programming," J. Soc. Indust. Appl. Math. *9*, 481–488 (1961).

[52] Bartels, R. H., G. H. Golub, and M. A. Saunders "Numerical Techniques in Mathematical Programming" in *Nonlinear Programming* (ed. by J. B. Rosen, O. L. Mangasavian and K. Ritter), Academic Press, New York, pp. 123–176 (1970).

MIN/MAX BOUNDS FOR DYNAMIC NETWORK FLOWS

W. L. Wilkinson

The George Washington University

ABSTRACT

This paper presents an algorithm for determining the upper and lower bounds for arc flows in a maximal dynamic flow solution. The procedure is basically an extended application of the Ford-Fulkerson dynamic flow algorithm which also solves the minimal cost flow problem. A simple example is included. The presence of bounded optimal arc flows entertains the notion that one can pick a particular solution which is preferable by secondary criteria.

I. INTRODUCTION

Ford and Fulkerson [1] introduced the notion of maximal dynamic flows in networks and provided an ingenious algorithm for solving the dynamic linear programming problem. A dynamic network consists of arcs and nodes with two nonnegative integers associated with each arc. One of the integers defines the capacity of the arc and the other the time required to traverse the arc. There are two distinguished nodes in the network, one for the source where all flows originate and one for the sink where all flows terminate. If at each node the commodity can either be transshipped immediately or held over for later shipment, what is the maximal amount of commodity flow from source to sink in a specified number of time periods? Solutions constructed by the Ford-Fulkerson algorithm have the attractive property of being presented as a relatively small number of activities (chain flows which represent a shipping schedule) which are repeated over and over (temporal repetition) until the end of the allotted time span. A consequence of this temporal repetition is that a single arc flow value in each arc represents an optimal solution independent of how these arc flows are decomposed into chain flows. In networks of operational interest, these optimal arc flow values frequently have an upper bound different from the lower bound. These bounds say that you can always find an optimal chain flow solution which lies on or between the stated range and no optimal solution lies outside these bounds. The procedure set forth in the sequel calculates these boundary values for each arc.

As shown in [2], the Ford-Fulkerson dynamic flow algorithm also solves the minimal cost flow problem. In this problem, roughly described, we are given a network having one or more sources and one or more sinks with availabilities at the sources and requirements at the sinks. There are intermediate nodes between the sources and sinks with connecting arcs having assigned capacities and unit shipping costs. The problem is to construct a feasible flow, if one exists, which minimizes cost in satisfying the requirements within the given availabilities. Similarly to the dynamic flow problem, the bounds on optimal arc flows will indicate the variety of ways, if any, in which such a feasible flow can be constructed.

Before describing the computing procedure for bounded flows, we will give a more formal statement of the dynamic flow problem referred to above.

II. MAXIMAL DYNAMIC FLOWS

Given the network $G=[N; A]$ with source s and sink t of the node set N, we let nonnegative integers $c(x, y)$ and $a(x, y)$ be the capacity and traversal time, respectively, of each arc (x, y) in the arc set A. Let $f(x, y; \tau)$ be the amount of flow that leaves x along (x, y) at time τ, consequently arriving at y at $\tau+a(x, y)$. Also $f(x, x; \tau)$ is the holdover at x from τ to $\tau+1$. If $V(P)$ is the net flow leaving s or entering t during the P periods 0 to 1, 1 to 2, . . ., $P-1$ to P, then the problem may be stated as the linear program:

$$\text{Maximize } V(P),$$

subject to

$$\sum_{\tau=0}^{P} \sum_{y} [f(s, y; \tau) - f(y, s; \tau - a(y, s))] - V(P) = 0,$$

$$\sum_{y} [f(x, y; \tau) - f(y, x; \tau - a(y, x))] = 0, \qquad x \neq s, t; \qquad \tau = 0, 1, \ldots, P,$$

$$\sum_{\tau=0}^{P} \sum_{y} [f(t, y; \tau) - f(y, t; \tau - a(y, t))] + V(P) = 0,$$

$$0 \leq f(x, y; \tau) \leq c(x, y).$$

Here $a(x, x) = 1$, $c(x, x) = \infty$ for holdovers at node x. If $f(x, y; \tau)$ and $V(P)$ satisfy the above constraints, we call f a dynamic flow from s to t (for P periods) and say that f has value $V(P)$. If also $V(P)$ is maximal, then f is a maximal dynamic flow.

III. BOUNDED FLOW ALGORITHM

To initiate the bounded arc flow computations, a maximal dynamic flow solution is required. The Ford-Fulkerson algorithm is used to obtain such a solution. We set forth their algorithm here, using the notation of [2], as Routine I in the interests of a coherent presentation for the convenience of the reader.

Routine II is an application of Kirchhoff's first general law on the conservation of flow at any node in a network. Routine II calculates "slack bounds"; slack in the sense that flow values contained by these bounds may not all be optimal, however, no optimal arc flow values are excluded by the bounds. Routine II is very efficient in calculating bounds based on local information at a node and is retained for that reason.

Routine III tightens up these bounds to their true value where necessary. Routine III is, essentially, an application of Routine I (Ford-Fulkerson algorithm) to a subnetwork composed of a special set of admissible arcs from the original network. Using the original flow solution and the flow boundaries from Routine II, optimal flows are circulated through the admissible arcs about a selected arc being scanned. Treating the initial and terminal nodes of the arc being scanned as a temporary source and sink, optimal flows are maximized and minimized in the arc thereby determining the absolute upper and lower optimal arc flow boundaries. If at any time the varying optimal flow values in the admissible

arcs are observed to reach a bound of Routine II, that bound has been verified since a known solution lies on a bound which is suspect of being too loose. Consequently, several unscanned arcs may get scanned in the process of scanning a particular arc. The reader may note that Routine III, slightly revised, could compute the true bounds without the aid of Routine II. Experience has shown that retaining the services of Routine II saves a substantial amount of computation time.

ROUTINE I

Ford-Fulkerson Algorithm

Initial Conditions

1. Establish P, the time span of interest.
2. Set node numbers $\pi(x)=0$ for all x.
3. Define $\bar{a}(x, y)=\pi(x)+a(x, y)-\pi(y)$ and consider as an admissible arc any (x, y) where $\bar{a}(x, y)=0$.
4. Set all $f(x, y)=0$.
5. During the routine a node is in one of the following states:
   Unlabeled and unscanned,
   Labeled but unscanned, or
   Labeled and scanned.
6. All nodes are unlabeled.

Arc Flow Generating Routine

STEP 1. To node s assign the label $[+t, \infty]$ and consider node s as unscanned.

STEP 2. Take any labeled, unscanned node x and suppose that it is labeled $[\pm w, h]$. To all nodes y that are unlabeled and such that:

a. (x, y) is admissible and $f(x, y) < c(x, y)$, assign the label $[+x, \min (h, c(x, y)-f(x, y))]$, or if

b. (y, x) is admissible and $f(y, x) > 0$, assign the label $[-x, \min (h, f(y, x))]$.

Consider node x as scanned and any newly labeled y-nodes as unscanned. Repeat until:

a. node t is labeled (breakthrough), or

b. no new labels are possible and node t is unlabeled (non-breakthrough).

STEP 3. If breakthrough results and suppose node t is labeled $[+y, h]$, replace $f(y, t)$ by $f(y, t)+h$. Next turn attention to node y. In general, if y is labeled $[+x, m]$, replace $f(x, y)$ by $f(x, y)+h$, or if y is labeled $[-x, m]$, replace $f(y, x)$ by $f(y, x)-h$. Next turn attention to node x. Ultimately the node s is reached; at this point stop the replacement process. Starting with the new flows thus generated, discard the old labels and repeat the above Arc Flow Generating Routine until no new labels are possible and node t cannot be labeled. When this condition results, proceed with the following Non-Breakthrough Processing Routine.

Non-Breakthrough Processing

STEP 1. Calculate a value of δ as follows: Define

$$A_1 = \{(x, y) \mid x \epsilon X, y \epsilon \bar{X}, \bar{a}(x, y) > 0\},$$

$$A_2 = \{(x, y) \mid x \epsilon \bar{X},\ y \epsilon X,\ \bar{a}(x, y) < 0\},$$

where X is the subset of labeled nodes and $\bar{X}$ is the complementary subset of unlabeled nodes. Let

$$\delta_1 = \min_{(x,y)\epsilon A_1 \neq \phi} [\bar{a}(x, y)], \text{ or } P + 1 - \pi(t) \qquad \text{if } A_1 = \phi,$$

$$\delta_2 = \min_{(x,y)\epsilon A_2 \neq \phi} [-\bar{a}(x, y)], \text{ or } P + 1 - \pi(t) \qquad \text{if } A_2 = \phi.$$

Then

$$\delta = \min(\delta_1, \delta_2).$$

Now define for all x, new node numbers $\pi'(x)$ as

$$\pi'(x) = \begin{cases} \pi(x) & \text{if } x \text{ is labeled.} \\ \min [\pi(x) + \delta;\ \ \pi(x) + P + 1 - \pi(t)], & \text{if } x \text{ is unlabeled.} \end{cases}$$

After new node numbers have been assigned, consider $\pi'(x)$ as $\pi(x)$.

STEP 2. If $\pi(t) < P + 1$, return to the Arc Flow Generating Routine. If $\pi(t) = P + 1$, algorithm terminates.

ROUTINE II

Slack Arc Flow Bounds

Initial Conditions

1. A maximal dynamic flow solution has been obtained for a particular P. Retain $\bar{a}(x, y)$ and $f(x, y)$ for all (x, y). Retain $\pi(x)$ for all x.
2. Set for all (x, y):

$$G(x, y)/g(x, y) = \begin{cases} 0/0, & \text{if } \bar{a}(x, y) > 0, \\ c(x, y)/0, & \text{if } \bar{a}(x, y) = 0, \\ c(x, y)/c(x, y), & \text{if } \bar{a}(x, y) < 0. \end{cases}$$

3. Add an arc (t, s) and set $G(t, s) = g(t, s) = \sum_y f(s, y)$.
4. Order all nodes in increasing $\pi(x)$ sequence with no preference where equality exists.
5. All nodes are unscanned.
6. If $\bar{a}(x, y) = 0$, then (x, y) is an admissible arc.

Procedure

STEP 1. Take the lowest ordered, unscanned node x and to all admissible arcs calculate and insert into the arc record

$$G'(x, y) = \min [G(x, y); \quad \sum_i G(i, x) - \sum_j g(x, j) + g(x, y)], \tag{1a}$$

$$g'(x, y) = \max [g(x, y); \quad \sum_i g(i, x) - \sum_j G(x, j) + G(x, y)]. \tag{1b}$$

If $G'(x, y) < G(x, y)$ or $g'(x, y) > g(x, y)$, consider y as unscanned.* Now consider the newly assigned $G'(x, y)$ and $g'(x, y)$ as $G(x, y)$ and $g(x, y)$, respectively. When all admissible arcs have been examined, consider x as scanned, proceed to the next lowest ordered unscanned node and scan that node. Scan all nodes.

STEP 2. When all nodes have been scanned, then consider all nodes as unscanned. Take the highest ordered, unscanned node y and to all admissible arcs calculate and insert in the arc record

$$G'(x, y) = \min [G(x, y); \quad \sum_j G(y, j) - \sum_i g(i, y) + g(x, y)], \tag{2a}$$

$$g'(x, y) = \max [g(x, y); \quad \sum_j g(y, j) - \sum_i G(i, y) + G(x, y)]. \tag{2b}$$

If $G'(x, y) < G(x, y)$ or $g'(x, y) > g(x, y)$, consider x as unscanned. Now consider the newly assigned $G'(x, y)$ and $g'(x, y)$ as $G(x, y)$ and $g(x, y)$, respectively. When all admissible arcs have been examined, consider y as scanned, proceed to the next highest ordered unscanned node and scan that node. Scan all nodes.

STEP 3. When all nodes have been scanned, then consider all nodes as unscanned. Take the lowest ordered, unscanned node x and to all admissible arcs calculate and insert in the arc record the results of Equation (1). Consider x as scanned. If $G'(x, y) < G(x, y)$ or $g'(x, y) > g(x, y)$ consider y as unscanned. Proceed to the next lowest ordered, unscanned node and scan that node. Scan all nodes. This terminates Routine II. Go to Routine III.

NOTE: As described above, Routine II sweeps from source to sink, sink to source and then source to sink. Computational experience has indicated that three sweeps achieve the best compromise between best bounds and reasonable computing times. One could specify repetitive sweeps until a complete sweep had been made with no changes to $G(x, y)/g(x, y)$. Alternatively, premature termination is allowed at any point since we are only seeking approximations to the true $G(x, y)/g(x, y)$.

ROUTINE III

Taut Arc Flow Bounds

Initial Conditions

1. Label all arcs as follows:
   a. "Gg" if $f(x, y) = G(x, y) = g(x, y)$,
   b. "$G+$" if $f(x, y) = G(x, y) \neq g(x, y)$,
   c. "$+g$" if $f(x, y) = g(x, y) \neq G(x, y)$, or
   d. "$++$" if none of the above is true.

---

*The scan state of y may be either scanned or unscanned. If neither condition is met, then this state remains unchanged. This potential redundancy is necessary to accommodate the $a(x, y) = 0$ instances where $\pi(x) = \pi(y)$, thus both arcs, (x, y) and (y, x) may be admissible. The stated inequalities prevent looping.

2. Mark all arcs labeled "Gg" and consider them scanned.
3. An unmarked arc is an admissible arc.
4. All nodes are unlabeled and unscanned.
5. If all arcs are scanned, terminate the routine. Otherwise, go to the procedure below.

Procedure

STEP 1. Take any unscanned arc (x, y) and consider x as s' and y as t'.

a. If (s', t') is labeled "$+g$" go to STEP (2) below and omit STEP (3) below.

b. If (s', t') is labeled "$G+$" go to STEP (3) below.

c. If (s', t') is labeled "$++$" go to STEPS (2) and (3) below.

STEP 2. To node t' assign the label $[+s', G(s', t') - f(s', t')]$. Take a labeled unscanned node x, initially t' is the only such node, and suppose it is labeled $[\pm w, h]$, to all nodes y that are unlabeled and

a. (x, y) is admissible and $f(x, y) < G(x, y)$, assign the label $[+x, \min(h, G(x, y) - f(x, y))]$,

or if

b. (y, x) is admissible and $f(y, x) > g(y, x)$, assign the label $[-x, \min(h, f(y, x) - g(y, x))]$.

Consider node x as scanned. Repeat until node s' is labeled (breakthrough) or no new labels are possible and node s' is unlabeled (non-breakthrough). If breakthrough results and node s' is labeled $[+y, h]$, replace $f(y, s')$ by $f(y, s') + h$ or if node s' is labeled $[-y, h]$, replace $f(s', y)$ by $f(s', y) - h$. In either case, if the arc (y, s') or (s', y) was previously considered scanned, it remains scanned bearing the label "Gg." If not previously scanned, consider the following cases.

a. Node s' is labeled $[+y, h]$. If the new $f(y, s') = G(y, s')$ and the current label is "$+g$," relabel the arc "Gg" and consider it scanned; or if the current label is "$++$," relabel the arc "$G+$" and consider it unscanned. If the new $f(y, s') < G(y, s')$, the arc retains its current label and remains unscanned.

b. Node s' is labeled $[-y, h]$. If the new $f(s', y) = g(s', y)$ and the current label is "$G+$," relabel the arc "Gg" and consider it scanned; or if the current label is "$++$," relabel the arc "$+g$" and consider it unscanned. If the new $f(s'y) > g(s'y)$, the arc retains its current label and remains unscanned.

Next, turn attention to node y and repeat the replacement and labeling process as for (y, s') or (s', y), incrementing or decrementing the flow value by h and determining whether the arc is to be considered scanned or unscanned. Continue this replacement process until a reverse path to node s' has been traced out. At this point, stop the replacement process. If $f(s', t') = G(s', t')$ a condition for non-breakthrough exists. If not, then starting with the new flows thus generated, discard the old node labels, consider all nodes as unscanned and repeat this Procedure until non-breakthrough results. When non-breakthrough results, record $G(s', t') = f(s', t')$. Erase all node labels, consider all nodes as unscanned and proceed to STEP 3 below if 1b is satisfied. Otherwise, go to STEP 4.

STEP 3. Do STEP 2 except assign the initial label to node s' as $[-t', f(s', t') - g(s', t')]$, stop the labeling and replacement breakthrough process at t' and, on nonbreakthrough, record $g(s', t') = f(s', t')$. Then go to STEP 4 below.

STEP 4. Consider (s', t') as scanned and reverted to its original (x, y) designation. Erase all node labels and consider all nodes as unscanned. Take the next unscanned arc (x, y) and repeat all of the above Procedure. If no such arc exists, terminate the routine. $G(x, y)$ and $g(x, y)$ are the firm upper and lower bounds, respectively, for optimal arc flows.

IV. PROOF OF EXTREME ARC FLOW VALUES

We stated earlier that the Bounded Flow Routine II has proven useful in reducing computing times. Since Routine III operates on the slack bounds produced by Routine II, it is necessary to show that optimal arc flow values are never excluded by Routine II. We will now prepare the way for stating such a theorem.

LEMMA 1: The sequence of equation pairs for $G'(x, y)$ and $g'(x, y)$ in the Bounded Flow Routine II will produce monotonic nonincreasing values of $G'(x, y)$ and monotonic nondecreasing values of $g'(x, y)$ for all (x, y).

PROOF: The truth of the assertion is a natural consequence of the structure of Equations (1) and (2) for $G'(x, y)$ and $g'(x, y)$ where the upper limit for $G'(x, y)$ is $G(x, y)$ and the lower bound for $g'(x, y)$ is $g(x, y)$. Independent of the number of iterations of the equation pairs, the previously calculated value for $G(x, y)$ or $g(x, y)$ provides a bound consistent with the assertion above.

LEMMA 2: If $G(x, y) \geq g(x, y)$ for all (x, y), Equations (1) and (2) will maintain $G'(x, y) \geq g'(x, y)$ for all (x, y).

PROOF: Consider the pair of Equations (1) for $G'(x, y)$ and $g'(x, y)$. The cases of interest are where:

$$G'(x, y) = \sum_i G(i, x) - \sum_j g(x, j) + g(x, y) < G(x, y), \tag{1}$$

and

$$g'(x, y) = \sum_i g(i, x) - \sum_j G(x, j) + G(x, y) > g(x, y). \tag{2}$$

Ignoring the inequalities and subtracting the second equation from the first, we have

$$G'(x, y) - g'(x, y) = \sum_i [G(i, x) - g(i, x)] + \sum_{j \neq y} [G(x, j) - g(x, j)].$$

By hypothesis, each of the two summation terms on the right side is nonnegative, therefore, the difference on the left side is nonnegative.

Equations (2) are symmetric with (1) for calculations progressing from sink to source. A similar exercise to the above would produce equivalent results.

That the hypothesis is true at the outset can be seen in the Initial Conditions established by Condition No. 2. Recall that we insisted that $c(x, y) \geq 0$.

LEMMA 3: When scanning a node x, the equations of Routine II always calculate $\sum_j G(x, j) \geq \sum_i g(i, x)$.

PROOF: Consider Equation (1b) which is

$$g'(x, y) = \max [g(x, y); \quad \sum_i g(i, x) - \sum_j G(x, j) + G(x, y)],$$

or

$$g'(x, y) - G(x, y) \geq \sum_i g(i, x) - \sum_j G(x, j).$$

We restate the above inequality by changing signs. Now,

$$G(x, y) - g'(x, y) \leq \sum_j G(x, j) - \sum_i g(i, x).$$

Lemma 2 states that $G'(x, y) \geq g'(x, y)$ and Lemma 1 assures us that $G(x, y) \geq G'(x, y)$. Therefore, the left side is nonnegative. Consequently, $\sum_j G(x, j) \geq \sum_i g(i, x)$.

Analogous arguments to those above for Lemma 3 would prove its corollaries for the following:

$\sum_i G(i, x) \geq \sum_j g(x, j)$ by considering Equation (1a),

$\sum_j G(y, j) \geq \sum_i g(i, y)$ by considering Equation (2a), and

$\sum_i G(i, y) \geq \sum_j g(y, j)$ by considering Equation (2b).

THEOREM 1: Let $G(x, y)$ and $g(x, y)$ be the upper and lower bounds, respectively, as determined by the Bounded Flow Routine II for all arcs (x, y) in $G = [N; A]$. Then any temporally repeated maximal dynamic flow solution for this network will contain for all (x, y) a flow $f(x, y)$ such that

$$g(x, y) \leq f(x, y) \leq G(x, y).$$

PROOF: Required flows are initially set in the inadmissible arcs according to the optimality criteria, i.e., for those arcs where $\bar{a}(x, y) < 0$, the minimum flow is equal to the arc capacity and for those arcs where $\bar{a}(x, y) > 0$, the maximum flow is equal to zero. Where $\bar{a}(x, y) = 0$, the flow may be anywhere between zero and the arc capacity. In the return arc from sink to source, we have set the minimum and maximum equal to the maximal static flow for the system for the time period of interest. These values do not change at any time. Initially for the admissible arcs the maximum flow is set at the arc capacity and minimum flow at zero which are the broadest possible bounds since we insist that $0 \leq f(x, y) \leq c(x, y)$ for all (x, y). Clearly then, our theorem holds for the initial conditions.

Lemma 1 tells us that $G(x, y)$ never increases, but tends to decrease and $g(x, y)$ never decreases, but tends to increase while Lemma 2 maintains that $G(x, y) \geq g(x, y)$ for all (x, y) at all times. Lemma 3 and its corollaries insure that the available flow into a node is at least equal to the required flow out of a node and, conversely, the required flow into a node is no greater than the available flow out of a node for all nodes at all times. Since initially the feasibility of meeting local conditions of optimality exists everywhere, we will proceed by induction on the sequence of nodes to be scanned, i.e.,

$$s, x_1, x_2, \ldots, t, \text{ where } \pi(x_i) \leq \pi(x_{i+1}).$$

Where equality holds between two or more nodes and the connecting arcs have zero traversal times, there may be redundant scanning of any of the nodes, but Lemmas 1 through 3 will hold nonetheless for each scan.

For the initial scanning of s, we know that $G(s, x)$ and $g(s, x)$ are all valid in the sense of the theorem. Therefore the cases of interest are where $G'(s, x) < G(s, x)$ and $g'(s, x) > g(s, x)$. Accordingly,

$$G'(s,x) = G(t,s) - \sum_{j \neq x} g(s,j),$$

or

$$G'(s,x) + \sum_{j \neq x} g(s,j) = G(t,s).$$

Suppose we substitute

$$f(s, x) > G'(s, x) \qquad \text{for } G'(s, x).$$

then

$$f(s, x) + \sum_{j \neq x} g(s, j) > G(t, s),$$

a contradiction to conservation of flow at s and the relative flow conditions maintained by Lemma 3.

Similarly,

$$g'(s,x) + \sum_{j \neq x} G(s,j) = g(t,s),$$

and we substitute $f(s, x) < g'(s, x)$ for $g'(s, x)$. Then

$$f(s, x) + \sum_{j \neq x} G(s, j) < g(t, s),$$

again a contradiction as above. Thus we see that $G'(s, x)$ and $g'(s, x)$ are valid upper and lower bounds, respectively, for all x, whether or not they have been changed from their initial values.

Next consider x_1. If $\sum_i G(i, x_1)$ or $\sum_i g(i, x_1)$ are no longer their original values, we know from the above that they are valid. Again, the equation of interest is

$$G'(x_1, y) + \sum_{j \neq y} g(x_1, j) = \sum_i G(i, x_1),$$

and a substitution $f(x_1, y) > G'(x_1, y)$ constitutes a contradiction. Similarly,

$$g'(x_1, y) + \sum_{j \neq y} G(x_1, j) = \sum_i g(i, x_1).$$

A substitution $f(x_1, y) < g'(x_1, y)$ produces a contradiction and $G'(x_1, y)$ and $g'(x_1, y)$ are validated.

As we proceed, we see that this holds for any x_j for the calculations are based on $\sum_i G(i, x_j)$ and $\sum_i g(i, x_j)$ which, although possibly changed from their initial values, are known to be valid in the sense of the theorem.

At the pivot node t, we reverse our sequence and proceed to s. Lemmas 2 and 3 insure that ultra-conservative conditions hold at this crucial point and a parallel argument to the s to t sequence would produce equivalent results. One may make as many iterative passes from s to t and t to s as desired without violating the conditions asserted by the theorem. This concludes our proof.

We now turn our attention to the Bounded Flow Routine III. The scanning process of Routine III is basically an application of the Ford-Fulkerson (F-F) algorithm (Routine I) as it operates between nonbreakthroughs. The maintenance of conservation of flow at every node and the maximizing properties of this algorithm, are well established [2]. Our proof for Routine III then reduces to showing that there exists a formal equivalence between our application and the standard conditions for the F-F algorithm.

Consider STEP 2 which is concerned with determining the best $G(x, y)$. $G(x, y)$ has replaced $c(x, y)$ as the upper bound and $g(x, y)$ has replaced zero as the lower bound. Justification for these substitutions can be found in Theorem 1. Our new network has all the arcs and nodes of the old except the arc being scanned. As in the F-F algorithm, only those arcs where $\bar{a}(x, y) = 0$ are admissible for labeling purposes because under the optimality criteria, it is only in these arcs that the flow can be altered. We start with a feasible flow as does the F-F algorithm. The source for this network is t' and the sink is s'. The source gets labeled with the maximum amount of flow that can be augmented in (s', t'), i.e., $G(x, y) - f(x, y)$ by Theorem 1. In the F-F algorithm, this is taken to be ∞ for it is not known *a priori* what the maximum amount of flow augmentation is. The labeling rules are the same as are the rules for breakthrough and nonbreakthrough. There is a distinction in the replacement process where the flow is incremented or decremented in a sequence of arcs which go from sink to sink. However, the last arc in this sequence is the arc being scanned which is not a part of the current network. When the routine reaches nonbreakthrough, further flow augmentation is impossible and we have the maximum flow in the arc being scanned.

The argument for STEP 3 follows that for STEP 2 above. The source for this network is s' and the sink is t'. The source gets a negative label with the maximum amount the flow in (s', t') can be reduced, i.e., $f(s', t') - g(s', t')$ by Theorem 1. On nonbreakthrough, the flow in (s', t') will have been decremented to its absolute minimum with respect to optimality and we record the final $g(x, y) = f(x, y)$. We can now state the following theorem.

THEOREM 2: Let $G(x, y)$ and $g(x, y)$ be the upper and lower bounds, respectively, as determined by the Bounded Flow Routine III for all arcs (x, y) in $G = [N; A]$. Then the integers n, such that

$$g(x, y) \leq n \leq G(x, y)$$

provide an exhaustive set of valid arc flows for which there exists an integer, temporally repeated, maximal dynamic flow solution for $G[N; A]$.

V. AN EXAMPLE

In Figure 1 is shown a simple network and a dynamic flow solution for the stabilization time of $P = 15$. Following stabilization time, the static flow of 10 is repeated each time period and new solutions

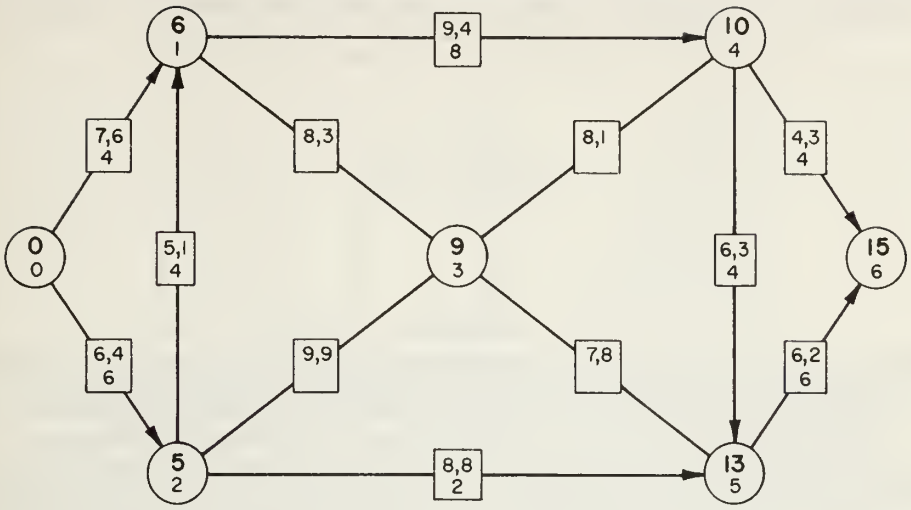

FIGURE 1.

are not necessary since the arc flows do not change in value. The small lower numbers in the nodes are the node names and the larger upper numbers are the node numbers $\pi(x)$'s. The data in the arc boxes are the following: Upper left—capacity; upper right—transit time; and the lower number is the arc flow $f(x, y)$. Capacities and transit times are symmetric, e.g., $c(x, y) = c(y, x)$.

In Figure 2 is shown the bounded arc flow solution based on the flow solution in Figure 1. The network data is the same as Figure 1 except the lower numbers in the arc boxes are the upper/lower bounds $G(x, y)/g(x, y)$.

In decomposing all alternative routes for their maximum optimal flow, we get the following set of nine routes.

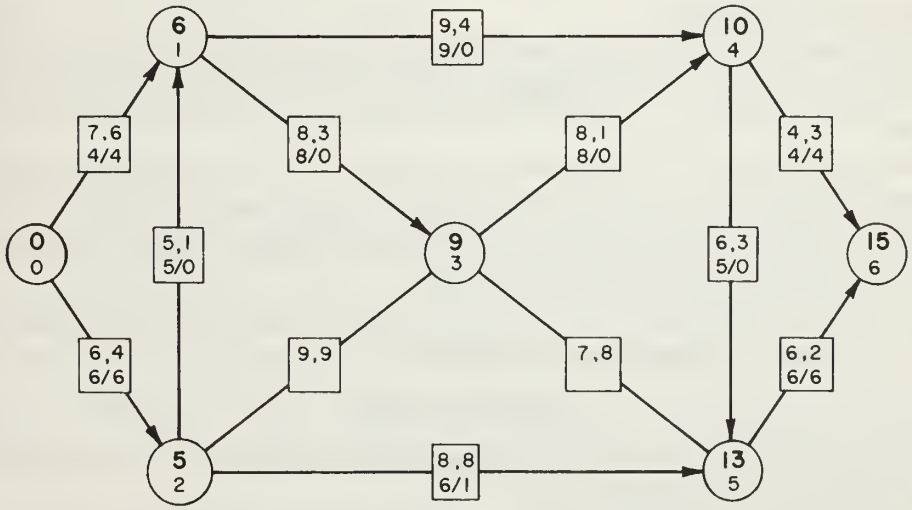

FIGURE 2.

Possible chain	Time length	$P=15$ use	Max flow
0–2–5–6	14	2	6
0–1–4–6	13	3	4
0–1–4–5–6	15	1	4
0–2–1–4–6	12	4	4
0–1–3–4–6	13	3	4
0–2–1–4–5–6	14	2	5
0–1–3–4–5–6	15	1	4
0–2–1–3–4–6	12	4	4
0–2–1–3–4–5–6	14	2	5

These alternative routes offer a fair variety of different ways in scheduling a particular optimal solution. For example, we list below two different solutions for contrast. Here, again, the time span is 15.

SOLUTION A

Chain	Flow	Use	Dynamic flow
0–1–4–6	4	3	12
0–2–5–6	6	2	12
			24

SOLUTION B

Chain	Flow	Use	Dynamic flow
0–1–4–5–6	4	1	4
0–2–1–4–6	4	4	16
0–2–5–6	2	2	4
			24

Consider, for instance, that Nodes 1 and 2 are origins and Nodes 4 and 5 are destinations. Then if there was some preference, not formally stated, for maximizing the origin-destination deliveries 1–4 and 2–5, one would choose Solution A. However, if the preferred pairings were 1–5 and 2–4, Solution B is the best.

VI. EXPERIENCE

Some version of the Bounded Flow Algorithm has been in use at The George Washington University since late 1967. The algorithm and the associated computer codes have been revised several times with the objective of increasing their efficiency. Currently, the bounded arc flow computation takes approximately one minute on a 500 arc network. The program is written in PL/1 for an IBM 360/50.

VII. ACKNOWLEDGEMENTS

The research was conducted as part of the Program in Logistics of the Institute for Management Science and Engineering, The George Washington University. The work was supported by the Office of Naval Research.

Special recognition is due to Donald J. Hunt of the Program in Logistics who, since the very beginning of this development, has made material contributions to the power and efficiency of the computational procedures. He is solely responsible for the large gains achieved in decreased running time which followed the original implementation of the algorithm.

Thanks are also due to Raymond W. Lewis of the Program in Logistics for his valuable observations and suggestions during the various levels of algorithmic development.

REFERENCES

[1] Ford, L. R., Jr. and D. R. Fulkerson, "Constructing Maximal Dynamic Flows from Static Flows," Operations Research *6*, 419–433 (1958).

[2] Ford, L. R., Jr., and D. R. Fulkerson, *Flows in Networks* (Princeton University Press, 1962).

PRODUCTION-ALLOCATION SCHEDULING AND CAPACITY EXPANSION USING NETWORK FLOWS UNDER UNCERTAINTY

Juan Prawda

Tulane University
New Orleans, Louisiana

ABSTRACT

This paper extends Connors and Zangwill's work in network flows under uncertainty to the convex costs case. In this paper the extended network flow under uncertainty algorithm is applied to compute N-period production and delivery schedules of a single commodity in a two-echelon production-inventory system with convex costs and low demand items. Given an initial production capacity for N periods, the optimal production and delivery schedules for the entire N periods are characterized by the flows through paths of minimal expected discounted cost in the network.

As a by-product of this algorithm the multi-period stochastic version of the parametric budget problem for the two-echelon production-inventory system is solved.

1. INTRODUCTION

In a recent paper Connors and Zangwill [7] developed the Network Flow Under Uncertainty (NFUU) or r-networks by allowing the requirements or availabilities at the nodes of a network to be discrete random variables with known probability distributions. They extended the standard deterministic multistage network flow problem introduced by Ford and Fulkerson [11]. The underlying structure of network flow problems was exploited in Ref. [7] to produce both a new structure which is not a deterministic network, but maintains many of its properties and a new node which replicates flows instead of conserving them. They called the former r-networks and the latter r-nodes. Construction of NFUU from a given N period stochastic problem is not given in this paper. The reader is referred to [7]. Given the N-period problem and convex objective criteria, we will develop an algorithm to calculate the network flow problem that minimizes expected cost.

Two applications of this algorithm are given:

1) To compute optimal N-period production and delivery schedules of a single commodity in a two-echelon production-inventory system with convex costs and low demand items; and,

2) To solve the parametric-budgetary problem in the multiperiod stochastic case, corresponding to the system described in (1).

This paper is organized as follows: In section 2 the Convex Network Flow Under Uncertainty Algorithm is stated, its validity and convergency proven; in section 3 the N-period, two-echelon production and delivery inventory problem is stated. In section 4 we extend the parametric budgetary problem to the multiperiod, stochastic case, for the system considered in section 3.

2. THE CONVEX NETWORK FLOW UNDER UNCERTAINTY ALGORITHM

Let $G=(\bar{N}, A)$ denote a Network Flow Under Uncertainty (NFUU), where $\bar{N}$ is a finite collection of elements $x, y, \ldots$ and A is a finite subset of ordered pairs (x, y) of elements taken from $\bar{N}$. $\bar{N}$ is

supposed to be of the form $\bar{N}=N_1 \cup N_2 \cup N_3$ with $N_i \cap N_j=\phi$ for $i, j=1, 2, 3, i \neq j$. The elements of N_1 are called nodes, the elements $r_1, r_2, \ldots$ of N_2 are called replication nodes or r-nodes and the elements $c_1, c_2, \ldots$ of N_3 are called collating nodes or c-nodes. Members of A are referred to as arcs. All arcs will be supposed to be of the form (x, y) with $x \neq y$, x, y in $\bar{N}$. We exclude arcs (x, y) where both x, y are in N_i, $i=2, 3$ and arcs going from r-nodes to c-nodes and vice versa.

If x is in $\bar{N}$, we let $a(x)$ ("after x") denote the set of all y in $\bar{N}$ for which (x, y) is in A, that is,

$$a(x)=\{y \epsilon \bar{N} \mid (x, y) \epsilon A\}.$$

Similarly, we let $b(x)$ ("before x") denote the set of all y in $\bar{N}$ for which (y, x) is in A, that is

$$b(x)=\{y \epsilon \bar{N} \mid (y, x) \epsilon A\}.$$

Given G, each arc (x, y) in A has associated with it a nonnegative real number $q(x, y)$, called the capacity of the arc (x, y) in A; a nonnegative integer $f(x, y)$ called the flow of the arc (x, y) in A; and a nonnegative real number $g(x, y)$, called the expected discounted cost of (x, y) in A. Both f and g are functions from A to the nonnegative reals, the former having nonnegative integers as its range. Let s, called the source, and t, called the sink, be two distinguished elements of $\bar{N}$.

Each r_k in N_2 has a single input arc and several output arcs and possesses the following two properties:

(1) $$f(x, r_k)=f(r_k, y) \qquad \text{for all } y \text{ in } a(r_k) \text{ and some } x \text{ in } b(r_k)$$

and

(2) $$g(x, r_k) \geq 0,\ g(r_k, y)=0 \qquad \text{for all } y \text{ in } a(r_k) \text{ and some } x \text{ in } b(r_k).$$

Property (1) merely states that flow on each of the output arcs of an r-node must be identical with that on the input arc, and (2) states that all the outgoing arcs of an r-node have an expected discounted cost of zero. Each c_k in N_3 is essentially the negative of an r-node, that is, it has several input arcs and a single output arc and possesses the following two properties:

(3) $$f(y, c_k)=f(c_k, x) \qquad \text{for all } y \text{ in } b(c_k) \text{ and some } x \text{ in } a(c_k)$$

and

(4) $$g(c_k, x) \geq 0,\ g(y, c_k)=0 \qquad \text{for all } y \text{ in } b(c_k) \text{ and some } x \text{ in } a(c_k).$$

Properties (3) and (4) are, respectively, the negative of (1) and (2).

It is shown in [7] that

(5) $$G=\bigcup_{i=1}^{M} G^i \cup N_2 \cup N_3,$$

where $G^i=(N_1^i, A^i)$, $i=1, \ldots, M$, M is a finite integer and $G^i \cap G^j=\phi$ for $i, j=1, \ldots, M$, $i \neq j$.

Each G^i, $i=1, \ldots, M$ is called a subnetwork and it is an ordinary network consisting of ordinary nodes x^i, y^i, . . . in N_1^i and ordinary arcs (x^i, y^i) in A^i where x^i, y^i are in N_1^i. In each subnetwork G^i the total inflow equals the total outflow, and the flow is conserved at each node in N_1^i, $(i=1, \ldots, M)$.

Two subnetworks, say G^i and G^j, $i, j=1, \ldots, M$, $i \neq j$, are connected if there exists at least one r_k in N_2 or c_k in N_3 for which (x^i, r_k) and (r_k, y^j) are in A when $x^i \epsilon N_1^i$ and $y^j \epsilon N_1^j$ or (x^j, c_k) and (c_k, y^i) are in A when $x^j \epsilon N_1^j$ and $y^i \epsilon N_1^i$. Let

$$N_2^i = \{r_k \epsilon N_2 \mid (r_k, x^i) \quad \text{or } (x^i, r_k) \epsilon A,\ x^i \epsilon N_1^i\}$$

$$N_3^i = \{c_k \epsilon N_3 \mid (c_k, x^i) \quad \text{or } (x^i, c_k) \epsilon A,\ x^i \epsilon N_1^i\}$$

be the sets of r and c nodes that connect subnetwork i $(i=1, \ldots, M)$ with the rest of the NFUU.

From the NFUU $G=(\bar{N}; A)$ (Figure 1) we observe that $N_1=\{1, 2, \ldots, 17, 18\}$, $N_2=\{r_1, r_2, r_3\}$, $N_3=\{c_1, c_2, c_3\}$, and $M=8$.

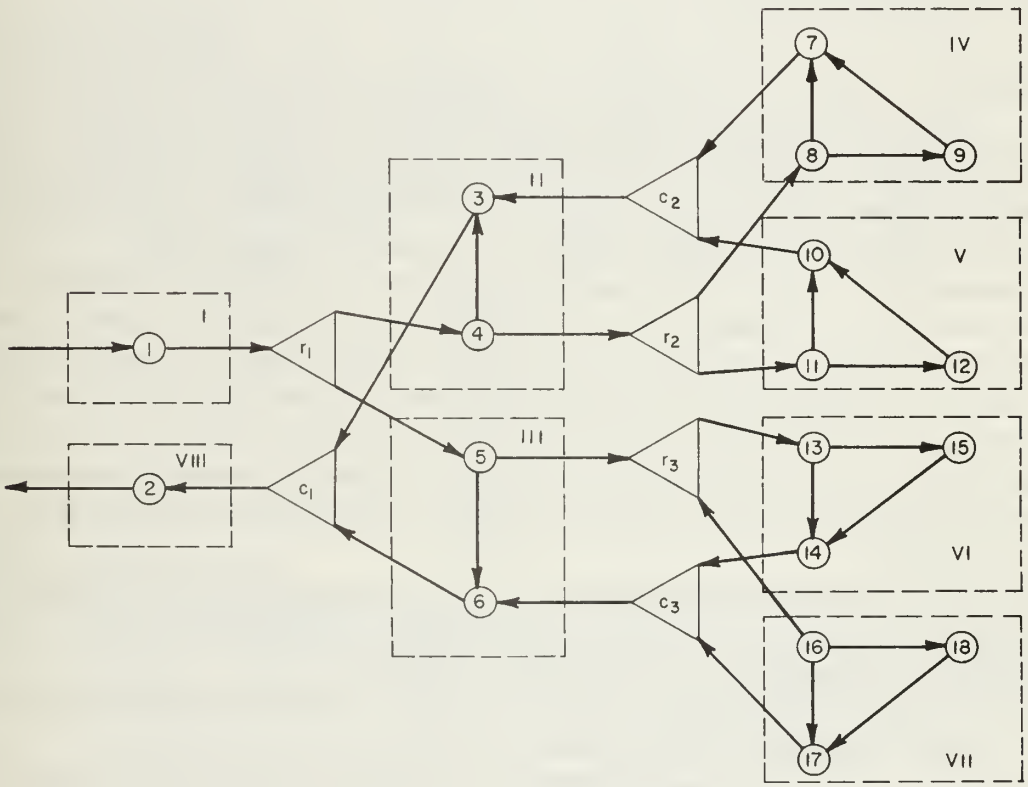

FIGURE 1. A network flow under uncertainty or r-network.

The algorithm to follow is based on the works of Connors and Zangwill [7] and Hu [14]. It is very closely related to the works of Beale [2], Busacker and Gowen [4], Hu [15], and Zangwill [24]. This algorithm, utilizing the decomposition (5), iterates by determining shortest routes or routes of minimal expected discounted cost along the subnetworks (which are ordinary networks) forcing one unit flow on this route and making appropriate adjustments for the r and c nodes.

Let $h[f(x, y)]$ be a nonnegative convex function on $f(x, y)$ for all (x, y) in A, such that $h(0) \equiv 0$ and the flow function $f(x, y)$ is required to have nonnegative integers as its range.

The cost function for the entire network is

$$\sum_{\text{all } (x, y) \in A} h[f(x, y)].$$

This cost function is a sum of convex functions and thus convex. Let

$$\bar{h}[f(x, y)] = h[f(x, y) + 1] - h[f(x, y)] \tag{6}$$

for $f(x, y) \geq 0$ and all (x, y) in A, and

$$\underline{h}[f(x, y)] = h[f(x, y) - 1] - h[f(x, y)] \tag{7}$$

for $f(x, y) > 0$ and all (x, y) in A.

Expression (6) defines the up-cost of an arc and (7) the down-cost of an arc. It is shown in [14], that $\bar{h}(\ ) > 0$ and $\underline{h}(\ ) < 0$; $\bar{h}(a) < \bar{h}(b)$ for $a < b$ and $|\underline{h}(a)| < |\underline{h}(b)|$. A particular flow called a *path-flow* is a flow with $f(s, x) = f(x, y) = \ . \ . \ . = f(z, t) = 1$ and $f(u, w) = 0$ for all $u, w \neq s, x, y, \ . \ . \ ., z, t$. If the cost of a flow with value v is known and we superimpose a path flow on this given flow, the resulting flow has value $v + 1$. $\bar{h}$ is used if the arc flow of the path flow is of the same direction as that of the arc flow of the flow with value v and $\underline{h}$ is used if the two flows are of opposite directions. The sum of $\bar{h}$ and $\underline{h}$ used in the path flow is called the *incremental cost* of the path flow.

An iteration of the convex NFUU algorithm first requires construction of a modified r-network from the current flow in the original r-network. A shortest route algorithm is then applied to determine the shortest route from source to sink in the modified r-network using the up cost and down cost ($\bar{h}(\ )$ and $\underline{h}(\ )$, respectively) of an arc as its length. One unit flow is then forced from source to sink in the original network along a route corresponding to the shortest path obtained in the modified r-network. The up cost and down cost of all the arcs in the modified r-network are redefined based on the new flow pattern of the original network. This cycle is repeated until the amount of flow at t in the original network is $\bar{I}$.*

Given G (the original network) and $\bar{I}$ (a positive integer corresponding to an input flow to the network), the precise algorithmic statements for the convex NFUU algorithm are:

STEP 0 (Initialization): Set $f(x, y) = 0$ for all (x, y) in A.

STEP 1 (Network Modification): Given the current flow in the original network $f(x, y)$ for all (x, y) in A, define a modified r-network as follows:

a) If $0 \leq f(x, y) \leq q(x, y)$ leave the arc (x, y) in the modified r-network with cost $\bar{h}[f(x, y)]$ as defined in (6)

b) If $f(x, y) = q(x, y)$, delete the arc (x, y) from the modified r-network

c) if $0 < f(x, y)$ and

i) x, y are in N_1, add a reverse arc (y, x) in the modified r-network with cost $\underline{h}[f(x, y)]$ as defined in (7)

ii) x is N_1 and y is N_2, add reverse arcs (y, x) and (z, y) in the modified r-network for all z in

*Definition of $\bar{I}$ is given in the next sentence.

$a(y)$, the former with cost $\underline{h}[f(x, y)]$ as defined in (7), and the latters with cost zero

iii) x is N_3 and y is N_1, add reverse arcs (y, x) and (x, z) in the modified r-network for all z in $b(x)$, the former with cost $\underline{h}[f(x, y)]$ as defined in (7) and the latters with cost zero.

Both b) and c) must be done if $f(x, y) = q(x, y)$ for all (x, y) in A.

STEP 2 (Shortest Route): Determine the shortest route or routes of minimal expected discounted cost from s to t in the modified r-network. Use properties (1) and (2) of r-nodes and (3) and (4) for c-nodes. Apply any shortest route algorithm [9, 11] with $\bar{h}(\ \ )$ and $\underline{h}(\ \ )$ as lengths in the arcs of the modified r-network.

STEP 3 (Flow Augmentation): Send one unit flow from s to t in the original network along the route corresponding to the shortest path just obtained in Step 2, that is, along the path whose incremental cost in the modified r-network relative to the existing flow in the original network is minimum.

STEP 4 (Iteration and Stopping Rule): If the amount at t in the original network is $\bar{I}$, stop. Otherwise return to Step 1 with the current flow.

If, during the application of Step 2, no shortest path exists from s to t in the modified r-network, the original problem is infeasible.

The validity and convergence of the algorithm is proven in [7] for the linear case. The next theorem proves the validity and convergence of the algorithm for the convex case.

Let $j=1$ stand for the source s and $j=m$ for the sink t. We will next prove that the convex NFUU algorithm is equivalent to compute a flow vector, $f=(f_i)$, $i=1, \ldots, k$ (k is the total number of elements in A) and a vector $b=(b_j)$, $j=2, \ldots, m-1$, whose components are the amounts of flow at the nodes ($b_1=\bar{I}$, $b_m=-\bar{I}$), which

$$\text{Min } h(f)$$

subject to

$$Df = b$$

(8)
$$If \leqslant q$$

$$f \geqslant 0,$$

where $Df=b$ merely states that:

a) the total inflow and total outflow of every node in N_1 must equal the amount of flow at the node,

b) every incoming flow (outgoing flow) of an r (c) node equals the amount of flow at the node,

c) every replication* (collating*) flow of an r (c) node equals the amount of flow at the node,

$q=(q_i)$, $i=1, \ldots, k$ is the vector of arc capacities, I is a $k \times k$ identity matrix and $h(f)=\sum_{i=1}^{k} [h_i(f_i)]$, where each $h_i(f_i)$ is a real-valued convex cost function on f_i, $i=1, \ldots, k$.

THEOREM 1: Assume that at the end of iteration $s(s > 0)$ of the convex NFUU algorithm, $\bar{f}^s$ is a feasible solution to the convex programming problem

(9)
$$\text{Min } h(f)$$

*The flow on each of the outgoing (incoming) arcs of an $r(c)$ node is called replication (collating) flow.

subject to

$$Df = b'$$

$$If \leq q$$

$$f \geq 0.$$

We let $\bar{f}^0 = 0$ and $\bar{f}^s < \bar{I}$ where $\bar{I}$ is the input flow to the NFUU.

For this $\bar{f}^s$ suppose $\hat{f}^s$ is optimal to

(10) $$\text{Min } h(f)$$

subject to

$$Df = \eta$$

$$-I\bar{f}^s \leq If$$

$$If \leq q - \bar{f}^s,$$

where

$$\hat{f}^s = (\hat{f}_i^s) = \begin{cases} 0 \text{ if arc } i \text{ is not in the path flow} \\ \\ 1 \text{ if arc } i \text{ is in the path flow,} \end{cases}$$

for $i = 1, \ldots, k$ and

$$\eta = (\eta_j) = \begin{cases} 0 \text{ if node } j \text{ is not in the path flow} \\ \\ 1 \text{ if node } j \text{ is in the path flow,} \end{cases}$$

for $j = 1, \ldots, m$. Then $f^{s+1} = \bar{f}^s + \hat{f}^s$ is optimal for

(11) $$\text{Min } h(f)$$

subject to

$$Df = b' + \eta$$

$$If \leq q$$

$$f \geq 0.$$

Pf: For $h(f)$ linear, the proof is in [7]. Assume $h(f)$ convex with $h(0) \equiv 0$. First we will prove f^{s+1} is feasible for (11). Adding the first m constraints of (9) and (10) yields

$$D(\bar{f}^s + \hat{f}^s) = b' + \eta$$

or

$$Df^{s+1} = b' + \eta.$$

f^{s+1} is always bounded below by zero, since $\bar{f}^s \geq 0$ from (9) and $\hat{f}^s \geq -\bar{f}^s$ from (10). The last $2k$ constraints of (10) yield

$$-I\bar{f}^s \leq I\hat{f}^s \leq q - \bar{f}^s,$$

adding $\bar{f}^s \geq 0$ to the above inequality

$$-I\bar{f}^s + I\bar{f}^s \leq I\hat{f}^s + I\bar{f}^s \leq q - \bar{f}^s + \bar{f}^s.$$

The above inequality yields

$$0 \leq If^{s+1} \leq q,$$

and thus f^{s+1} is feasible to (11).

Next, we prove the optimality of f^{s+1}. Let the cost associated with problem (9) be $h(\bar{f}^s)$ and let $h(\hat{f}^s)$ be the optimal incremental cost of problem (10) corresponding to the optimal path flow $\hat{f}^s$. It follows from the optimality of $\hat{f}^s$ that $h(\hat{f}^s) \leq h(f^s)$ for any path flow f^s feasible to problem (10).

Since the cost of the new flow f^{s+1} in the original network equals the cost of the old existing flow $\bar{f}^s$ in the original network plus the incremental cost of the path flow $\hat{f}^s$, in the modified r-network, it follows that

$$h(f^{s+1}) = h(\bar{f}^s + \hat{f}^s) = h(\bar{f}^s) + h(\hat{f}^s) \leq h(\bar{f}^s) + h(f^s) = h(\bar{f}^s + f^s) = h(f^*)$$

for any feasible f^* to problem (11). Then f^{s+1} is optimal to (11). Q.E.D.

The last theorem proves that in terms of the NFUU, at the end of each iteration, the flow will be optimal for the amount thus far placed into the source of the NFUU. The convergency of the algorithm follows from the fact that f^{s+1} is bounded above by a finite integer $\bar{I}$ and at each iteration f^{s+1} increases by a flow of value one.

Next we suggest the application of the above convex NFUU algorithm to the solution of two problems, one given in section 3 and the other one in section 4.

3. AN N-PERIOD, 2-ECHELON PRODUCTION-INVENTORY SYSTEM

Interest in the multiechelon inventory systems has been spurred by the existence of large military logistics networks and private industry. There are a number of papers on single and multiproduct, multiinstallation inventory models that have been published. A comprehensive review of these topics can be found in the excellent published bibliographies of Iglehart [17], [18], Scarf, Gilford and Shelley [20], and Veinott [22].

Several approaches have been used to compute optimal N-period reordering points in the preceding multiechelon inventory systems. For instance, Bessler and Veinott [3], using the assumption that stock left over (backlogged) at the end of the N periods in each facility can be salvaged (purchased) at the same stationary unit price, decompose an N variable linear cost function into the sum of $N-$one variable linear functions, and a stationary policy given by a critical vector is shown to be optimal.

Relaxing Bessler and Veinott's [3] assumption, the N-period problem will, in general, not decompose to $N-$one period problems and dynamic programming is used to compute the optimal policy. Others, such as Clark and Scarf [5, 6] have used dynamic programming. However, its use has been shown to be computationally infeasible for even simpler problems than the one considered by Bessler and Veinott. (See [17, 18].)

The objective of this section is to suggest the use of the previous convex NFUU algorithm to solve N-period, multiechelon production and delivery inventory systems. It is the structure of the NFUU networks that allows for some computational improvement to obtain optimal production and delivery schedules of N-period, multiechelon stochastic inventory problems, with low item demands, with respect to other techniques used to solve similar systems, such as dynamic programming.

This paper is concerned with the problem of scheduling the production $x_{01}, x_{02}, \ldots, x_{0N}$ and allocation $x_{11}, \ldots, x_{1N}, x_{21}, \ldots, x_{2N}, \ldots, x_{n1}, \ldots, x_{nN}$ of a single product in facilities $1, \ldots, n$ in successive time periods $1, 2, \ldots, N$ so as to minimize the total expected discounted costs over the N periods. The requirements of each facility are discrete random variables each of which has known probability mass function.

Figure 2 illustrates a two-echelon system consisting of a plant, a warehouse 0, and n facilities numbered $1, 2, \ldots, n$. Although we are interested in more general multiechelon systems, the preceding one will suffice to illustrate our approach. Some remarks concerning the generalization to more complex multiechelon systems are given at the end of this section.

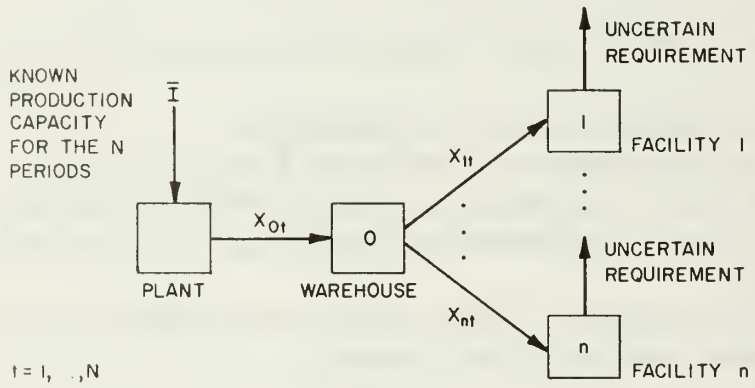

FIGURE 2.

At the beginning of period one we consider a known production capacity for the N periods. Let $\bar{I}$ denote the production capacity. The production capacity could be present in this problem due to restrictions of raw material to produce the given single commodity. Uncertain requirements in facility i $(i=1, \ldots, n)$ in each period are satisfied insofar as possible from stock on hand at the beginning of the period in that facility and from the allocation and production of the commodity at the beginning of the period. Requirements which cannot be met in a given period (because, for example, limited production) are backlogged until they can be satisfied by subsequent production or allocation in future periods.

Let α_{it} $(i=1, \ldots, n; t=1, \ldots, N)$ be a parameter associated with each facility i in any period t. this parameter will take the following values:

$$\alpha_{it}=1,2, \ldots, n_{it} \qquad \text{for } i=1, \ldots, n \text{ and } t=1, \ldots, N.$$

Let $\{D_{it}, i=1, \ldots, n; t=1, \ldots, N\}$ be a family of discrete, nonnegative random variables. For a fixed i and t, D_{it} takes on values in $\{d_1, d_2, \ldots, d_{n_{it}}\}$, a set of nonnegative real numbers. Let

$$P_{\alpha_{it}}=P\{D_{it}=d_{\alpha_{it}}\} \qquad \text{for all } \alpha_{it}.$$

The sequence $\{P_{\alpha_{it}}\}$ of real numbers will be a probability distribution of D_{it} with $P_{\alpha_{it}} \geq 0$, and

$$\sum_{\alpha_{it}=1}^{n_{it}} P_{\alpha_{it}}=1.$$

This distribution is assumed to be known. D_{it} is not necessarily independent* between any two facilities or identically distributed for successive periods. Let $\omega_{it} \underline{\underline{\Delta}} (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{it})$, $i=1, \ldots, n$ and $1 \leq t \leq N$, be the index associated with the random variables defined below. This notation identifies the sequence of realizations in facility i up to period t. Thus for example $\omega_{i3} \underline{\underline{\Delta}} (2, 1, 5)$ denotes the sequence of following events in facility i: In period 1 realization 2, in period 2 realization 1, and realization 5 in period 3.

Let

$$\omega_{0t} \underline{\underline{\Delta}} (\omega_{11}, \omega_{21}, \ldots, \omega_{n1}, \ldots, \omega_{1t}, \omega_{2t}, \ldots, \omega_{nt}).$$

We let $D_{it}^{(\omega it)}$ for $i=1, \ldots, n$ and $t=1, \ldots, N$ be the vector of realizations caused by the stochastic requirements in facility i up to period t. It is conditioned on all previous realizations in that facility in previous time periods, that is $D_{i1}^{(\omega i1)}, \ldots, D_{i,t-1}^{(\omega i, t-1)}$ and occurs with conditional probability $\underline{P}(\omega it)$, where

$$\underline{P}(\omega_{it})=\underline{P}\{D_{it}=d_{\alpha_{it}} \mid D_{i,t-1}=d_{\alpha_{i,t-1}}, \ldots, D_{i1}=d_{\alpha_{i,1}}\}.$$

Let $x_{0t}^{(\omega 0t)}$ (≥ 0) be the production completed in the plant (warehouse) at the beginning of period t given the sequence of random events ω_{0t} and $x_{it}^{(\omega it)}$ (≥ 0) $(i=1, \ldots, n)$, the allocation completed in facility i at the beginning of period t $(t=1, \ldots, N)$ given the sequence of random events ωit. Let $I_{0t}^{(\omega 0t)}(t=1, \ldots, N)$ be the inventory at the end of period t in the warehouse given the sequence ω_{0t} and $I_{it}^{(\omega it)}(i=1, \ldots, n; t=1, \ldots, N)$ be the inventory at the end of period t in each facility i given the sequence ω_{it}. We will assume that the initial inventory at all facilities and the warehouse is zero at the beginning of period one.

In order to simplify the statement of the problem, we are provisionally going to suppress the index ω_{it} and merely refer to the random variables x_{it} and I_{it}, $i=0, 1, \ldots, n$, $t=1, \ldots, N$.

---

*If for a fixed i, $D_{it}(t=1, \ldots, N)$ is a sequence of dependent random variables, then the marginal distribution $P_{\alpha_{ik}}$ $(1 \leq k \leq N)$ can be obtained from the given joint probability distribution.

We will assume that the lead time in production and delivery to the n facilities is zero. The inventory level equation becomes

$$I_{it}=\begin{cases}\sum_{h=1}^{t}(x_{ih}-\sum_{k=1}^{n}x_{kh}) & \text{for } i=0\\ \sum_{h=1}^{t}(x_{ih}-D_{ih}) & \text{for } i=1,\ \ldots,\ n,\end{cases}$$

where $t=1, \ldots, N$. Let β_i be a nonnegative integer denoting the number of periods of backlog permitted for facility i, thus

$$I_{it}\geq-\sum_{k=t-\beta_i+1}^{t}D_{ik} \qquad \text{for all } i=1,\ \ldots,\ n.$$

Note that, in general, we would have

$$\sum_{t=1}^{N}x_{0t}\leq\bar{I},^{*}$$

which implies that it may be optimal not to use all the production capacity $\bar{I}$ through the N periods. Let k_{it} be the known capacity of the production line for $i=0$ and of the transportation facilities for $i>0$ $(i=1, \ldots, n)$ during period t $(t=1, \ldots, N)$. Let Q_{it} be the known storage capacity for the warehouse $(i=0)$ and the n facilities $(i=1, \ldots, n)$ during period t $(t=1, \ldots, N)$. Thus we will require that

$$x_{it}\leq k_{it} \text{ for } i=0, \ldots, n \text{ and } t=1, \ldots, N$$

and

$$I_{it}\leq Q_{it} \qquad \text{for } i=0, \ldots, n \text{ and } t=1, \ldots, N.$$

Let

$$z=\left(x_{01}^{(\omega_{01})}, \ldots, x_{0N}^{(\omega_{0N})}, x_{11}^{(\omega_{11})}, \ldots, x_{1N}^{(\omega_{1N})}, \ldots, x_{n1}^{(\omega_{n1})}, \ldots, x_{nN}^{(\omega_{nN})}\right)$$

be the schedule vector for the entire system, given the sequence of realizations in all facilities i, $i=1, \ldots, n$ up to period N.

We have the following costs in each period: production, shipping, holding, and shortage (the last one due to backlogged demand). The preceding costs are assumed to be convex functions of the quantities produced and delivered at the beginning of the period and of the quantities stored or backlogged

*In terms of the NFUU this can be accomplished by arcs representing unused production capacity in the plant at each period t, $t=1, 2, \ldots, N$.

at the end of the period, respectively. The cost functions for successive periods need not be the same.

Let δ_t, $0 \leq \delta_t \leq 1$ be the discount factor for period t. Let $\gamma_1 = 1$ and $\gamma_t = \prod_{j=1}^{t-1} \delta_t$, for $t > 1$. The total expected discounted cost $F(z)$ is defined to be the sum of the following expected discounted costs:

i) Total expected discounted production and holding cost in the warehouse:

$$\sum_{t=1}^{N} \gamma_t \cdot [C_t(x_{0t}) + H_{0t}(I_{0t})].$$

ii) Total expected discounted transportation, holding, and penalty cost in all facilities:

$$\sum_{t=1}^{N} \sum_{i=1}^{n} \gamma_t \cdot [T_{it}(x_{it}) + H_{it}(I_{it})],$$

where

$$H_{it} = \text{Max}\,\{h_{it}(I_{it}),\, p_{it}(I_{it})\},$$

where $h_{it}(\cdot)$, $p_{it}(\cdot)$ are convex functions of their arguments satisfying

$$h_{it}(I_{it}) - p_{it}(I_{it}) = \begin{cases} > 0 \text{ if } I_{it} > 0 \\ 0 \text{ if } I_{it} = 0 \\ < 0 \text{ if } I_{it} < 0, \end{cases}$$

$h_{it}(\cdot)$ and $p_{it}(\cdot)$ are, respectively, the expected holding and penalty costs,* for $i = 0, \ldots, n$ and $t = 1, \ldots, N$. $C_t(\cdot)$, $h_{it}(\cdot)$, $p_{it}(\cdot)$, and $T_{it}(\cdot)$ are convex functions of their respective arguments with $C_t(0) = h_{it}(0) = p_{it}(0) = T_{it}(0) \equiv 0$ for $i = 0, \ldots, n$ and $t = 1, \ldots, N$.

Then the problem can be stated as: given the uncertain market requirements for each of the n facilities over the next N periods and the production capacity for the N periods, find a production-allocation schedule z, called optimal, which minimizes the total expected discounted cost

$$F(z) = \sum_{t=1}^{N} \gamma_t \cdot [C_t(x_{0t}) + H_{0t}(I_{0t}) + \sum_{i=1}^{n} \{T_{it}(x_{it}) + H_{it}(I_{it})\}]$$

subject to

$$\sum_{t=1}^{N} x_{0t} \leq \bar{I},$$

---

*$H_{it}(I_{it})$ may not be convex for $I_{it} = 0$. However, in terms of the network flow approach there are going to be arcs with expected cost $h_{it}(\cdot)$ corresponding to storage of inventory in facility i at the end of period t and different arcs with expected cost $p_{it}(\cdot)$ corresponding to backlogged inventory in facility i at the end of period t. Thus $H_{it}(I_{it})$ is never used explicitly in the NFUU.

$$0 \leq x_{it} \leq K_{it} \qquad \text{for } i=0, \cdot, \ldots, n \text{ and } t=1, \ldots, N,$$

and

$$I_{it} \leq Q_{it} \text{ for } i=0, \cdot, \ldots, n \text{ and } t=1, \ldots, N,$$

where

$$I_{it}=\begin{cases} \sum_{h=1}^{t}\left(x_{ih}-\sum_{k=1}^{n} x_{kh}\right) & \text{for } i=0 \\ \sum_{h=1}^{t}(x_{it}-D_{it}) & \text{for } i=1, \ldots, n \end{cases}$$

and

$$I_{it} \geq -\sum_{k=t-\beta_i+1}^{t} D_{ik}.$$

In order to solve the above problem we suggest that the above two-echelon multi-period, stochastic, production-delivery problem be rewritten in terms of the NFUU (see [7]), and then that the convex NFUU algorithm (described in section 2), be used to compute the optimal production and delivery schedules. Since the NFUU decomposes into a set of subproblems which are small network flow problems, this algorithm seems more attractive than dynamic programming, specially for multiechelon inventory systems with low demand items (0–1 demands) and small number of time periods. It was shown in Ref. [7] that the amount of computer storage required by the NFUU algorithm is proportional only to n, the dimension of the one-stage problem.

The approach suggested in this paper can be applied to more general multiechelon systems than the one depicted in Figure 2 (where transhipments between facilities are allowed), with consequent increase in the number of arcs and nodes in the NFUU.

4. A MULTIPERIOD, STOCHASTIC VERSION OF THE PARAMETRIC BUDGET PROBLEM

Suppose that a fixed budget of ν dollars can be allocated among the production-line, transportation, and storage facilities of the existing production-delivery inventory system for the purpose of increasing the production capacity $\bar{I}$ for the N periods. The cost of increasing the capacity of the production line $(i=0)$ and transportation facilities $(i=1, \ldots, n)$ during period $t(t=1, \ldots, N)$ is $\nu_{i,t}$ dollars per unit increase. The cost of increasing storage capacity at the warehouse $(i=0)$ and the n facilities $(i=1, \ldots, n)$ during period $t(t=1, \ldots, N)$ is $\nu_{s_{i,t}}$ dollars per unit increase.

Let $w_{it}^{(\omega_{it})}$ and $w_{s_{it}}^{(\omega_{it})}$ be decision variables corresponding, respectively, to the amount of additional capacity to be built in the production-line $(i=0)$, transportation $(i=1, \ldots, n)$, and storage facilities during period $t(t=1, \ldots, N)$ dependent upon the sequence ω_{it} of random events in facility i up to period t. Then the problem of increasing the production capacity for the N periods to, say, I' $(I'>\bar{I})$ is to minimize the following expected expansion cost.

$$\text{Min} \sum_{i=0}^{n} \sum_{t=1}^{N} (\nu_{it} \cdot \underline{P}(\omega_{it}) \cdot w_{it}^{(\omega_{it})} + \nu_{s_{it}} \cdot \underline{P}(\omega_{it}) \cdot w_{s_{it}}^{(\omega_{it})})$$

subject to

$$\sum_{t=1}^{N} x_{0t}^{(\omega_{0t})} = I'$$

$$0 \leq x_{it}^{(\omega_{it})} \leq K_{it} + w_{it}^{(\omega_{it})} \qquad \text{for } i=0, 1, \ldots, n \text{ and } t=1, \ldots, N$$

$$I_{it}^{(\omega_{it})} \leq Q_{it} + w_{s_{it}}^{(\omega_{it})} \qquad \text{for } i=0, \cdot, \ldots, n \text{ and } t=1, \ldots, N,$$

where

$$I_{it}^{(\omega_{it})} = \begin{cases} \sum_{h=1}^{t} (x_{ih}^{(\omega_{ih})} - \sum_{k=1}^{n} x_{kh}^{(\omega_{kh})}) & \text{for } i=0 \\ \sum_{h=1}^{t} (x_{it}^{(\omega_{it})} - D_{it}^{(\omega_{it})}) & \text{for } i=1, \ldots, n \end{cases}$$

$$I_{it}^{(\omega_{it})} \geq -\sum_{k=t-\beta_i+1}^{t} D_{ik}^{(\omega_{ik})}.$$

A related problem to the one just given above would be to maximize the production capacity I' (now a decision variable) for the N-periods with a fixed budget of ν dollars. This problem is stated as

$$\text{Max } I'$$

subject to

$$\sum_{i=0}^{n} \sum_{t=1}^{N} (\nu_{it} \cdot \underline{P}(\omega_{it}) \cdot w_{it}^{(\omega_{it})} + \nu_{s_{it}} \cdot \underline{P}(\omega_{it}) \cdot w_{s_{it}}^{(\omega_{it})}) = \nu$$

$$0 \leq x_{it}^{(\omega_{it})} \leq k_{it} + w_{it}^{(\omega_{it})} \qquad \text{for } i=0, \cdot, \ldots, n \text{ and } t=1, \ldots, N$$

$$I_{it}^{(\omega_{it})} \leq Q_{it} + w_{s_{it}}^{(\omega_{it})} \qquad \text{for } i=0, \ldots, n \text{ and } t=1, \ldots, N,$$

where

$$I_{it}^{(\omega_{it})} = \begin{cases} \sum_{h=1}^{t} (x_{ih}^{(\omega_{ih})} - \sum_{k=1}^{n} x_{kh}^{(\omega_{kh})}) & \text{for } i=0 \\ \sum_{h=1}^{t} (x_{it}^{(\omega_{it})} - D_{it}^{(\omega_{it})}) & \text{for } i=1, \ldots, n \end{cases}$$

$$I_{it}^{(\omega_{it})} \geq -\sum_{k=t-\beta_i+1}^{t} D_{ik}^{(\omega_{ik})}.$$

In terms of the NFUU $G=[\bar{N}; A]$ the preceding two problems are seen to be an extension of the deterministic parametric budget problem solved by Fulkerson [13] and Hu [14, 16], to the multiperiod, stochastic case.

The algorithm for solving the preceding two problems now follows:

STEP 0 (Initialization): Set $f(x, y)=0$ for all (x, y) in A.

STEP 1 (Network Modification): Given $f(x, y)$ for all (x, y) in A define a modified NFUU as follows:

a) If $f(x, y) < q(x, y)$ then $\bar{h}[f(x, y)]=0$;

b) If $f(x, y) \geq q(x, y)$ then $\bar{h}[f(x, y)]=g(x, y)$;

c) If $0 < f(x, y) \leq q(x, y)$ then $\underline{h}[f(x, y)]=0$;

d) If $f(x, y) > q(x, y)$ then $\underline{h}[f(x, y)]=-g(x, y)$;

where $g(x, y)$ is the capacity expansion unit cost for arc (x, y) in A and $q(x, y)$ is the original capacity of arc (x, y) in A. Obviously properties (1) and (2) of r-nodes and (3) and (4) of c-nodes must hold.

STEP 2 (Shortest Route): Send one unit flow from s to t in the original network along a route corresponding to the shortest path just calculated in the modified network, that is, the path in the modified network whose incremental cost is minimum. Apply any shortest route algorithm [9, 11] with $\bar{h}(\ )$ and $\underline{h}(\ )$ as lengths.

STEP 3 (Flow Augmentation and Stopping Rule): If the amount of flow at t is I' in the original network or the total amount of money used up is ν, stop; otherwise return to Step 1 with the current flow.

It is obvious that*:

$$w_{it} \text{ or } w_{s_{it}}=f(x, y)-q(x, y) \qquad \text{if } f(x, y) > q(x, y) \text{ for some } (x, y) \text{ in } A, \text{ and}$$

$$w_{it} \text{ or } w_{s_{it}}=0 \qquad \text{if } f(x, y) \leq q(x, y) \text{ for some } (x, y) \text{ in } A.$$

ACKNOWLEDGMENTS

My sincere thanks to Professors Gordon P. Wright and Larry R. Arnold for their helpful commentaries.

BIBLIOGRAPHY

[1] Arrow, K., S. Karlin, and H. Scarf, (eds.), *Studies in the Mathematical Theory of Inventory and Production* (Stanford University Press, Stanford Calif., 1958).

[2] Beale, E. M. L., "An Algorithm for Solving the Transportation Problem when the Shipping Cost over each Route is Convex," Nav. Res. Log. Quart. *6*, 43–56 (1959).

[3] Bessler, S. A. and A. F. Veinott, Jr., "Optimal Policy for a Dynamic Multi-Echelon Inventory Model," Nav. Res. Log. Quart. *13*, 335–389 (1966).

[4] Busacker, R. G. and P. J. Gowen, "A Procedure for Determining a Family of Minimal Cost Network Flow Patterns," Technical Rept. No. 15, Operations Research Office, Johns Hopkins University, Baltimore, Md. (1961).

*We assume that the function mapping the subscripts (i, t) $i=0, \cdot, :, \ldots, n$, $t=1, \ldots, N$ with the arcs (x, y) in A is known from the structure of the NFUU.

[5] Clark, A. and H. Scarf, "Optimal Policies for a Multi-Echelon Inventory Problem," Management Science, *6*, 475–490 (1960).

[6] Clark, A. and H. Scarf, "Approximate Solutions to a Simple Multi-Echelon Inventory Problem," Chapter 5 in *Studies in Applied Probability and Management Science* by Arrow, Karlin, and Scarf (eds.), Stanford, University Press, Stanford Calif. (1962).

[7] Connors, M. and W. Zangwill, "Cost Minimization in Networks with Discrete Stochastic Requirements," Operations Research *19*, 794–821 (1971).

[8] Dantzig, G., *Linear Programming and Extensions* (Princeton University Press, Princeton, N.J., 1963).

[9] Dreyfus, S. E., "An Appraisal of Some Shortest-Path Algorithms," Operations Research *17*, 395–412 (1969).

[10] El-Agizy, M., "Dynamic Inventory Models and Stochastic Programming," IBM Journal of Research and Development (1969), pp. 351–356.

[11] Ford, L. R. and D. R. Fulkerson, *Flows in Networks* (Princeton University Press, Princeton, N.J. 1963).

[12] Ford, L. R. and D. R. Fulkerson, "Constructing Maximal Dynamic Flows from Static Flows," Operations Research, *6*, 419–433 (1958).

[13] Fulkerson, D. R., "Increasing the Capacity of a Network, The Parametric Budget Problem," Management Science *5*, 472–483 (1959).

[14] Hu, T. C., "Minimum Convex Cost Flows," Nav. Res. Log. Quart. *13*, 1–19 (1966).

[15] Hu, T. C., "Recent Advances in Network Flows," SIAM Review *10*, 354–359 (1968).

[16] Hu, T. C., *Integer Programming and Network Flows* (Addison Wesley Publishing Co., Reading, Mass., 1969).

[17] Iglehart, D., "Recent Results in Inventory Theory," J. Indust. Eng. *18*, 48–51 (1967).

[18] Iglehart, D., "Recent Developments in Stochastic Inventory Models," Invited Paper at the National Meeting of ORSA, June 19, 1969, Denver, Colorado.

[19] Prawda, J. and G. P. Wright, "On Some Applications of Network Flows Under Uncertainty," Proceedings of the International IEEE Conference on Systems, Networks, and Computers, Oaxtepec, Morelos, Mexico (Jan. 19–21, 1971).

[20] Scarf, H. E., D. Gilford, M. Shelley, *Multistage Inventory Models and Techniques* (Stanford University Press, Stanford, Calif., 1963).

[21] Veinott, A., Jr., "Optimal Policy for Multiproduct, Dynamic, Nonstationary Inventory Problem," Management Science *12*, 206–222 (1965).

[22] Veinott, A., Jr., "The Status of Mathematical Inventory Theory," Management Science *12*, 745–777 (1966).

[23] Zangwill, W., "A Deterministic Multiproduct, Multi-facility, Production and Inventory Model," Operations Research *14*, 486–507 (1966).

[24] Zangwill, W., "The Shortest Route Problem under Either Concave or Convex Costs," Presented at the 12th Annual Operations Research Society of America Meeting, Santa Monica, California (1966).

[25] Zangwill, W., *Nonlinear Programming, A Unified Approach* (Prentice Hall, Inc., Englewood Cliffs, N.J., 1969).

CONCAVE MINIMIZATION OVER A CONVEX POLYHEDRON

Hamdy A. Taha

University of Arkansas

ABSTRACT

A general algorithm is developed for minimizing a well defined concave function over a convex polyhedron. The algorithm is basically a branch and bound technique which utilizes a special cutting plane procedure to identify the global minimum extreme point of the convex polyhedron. The indicated cutting plane method is based on Glover's general theory for constructing legitimate cuts to identify certain points in a given convex polyhedron. It is shown that the crux of the algorithm is the development of a linear underestimator for the constrained concave objective function. Applications of the algorithm to the fixed-charge problem, the separable concave programming problem, the quadratic problem, and the 0–1 mixed integer problem are discussed. Computer results for the fixed-charge problem are also presented.

I. INTRODUCTION

Consider the problem

$$\min_{x \in Q} f(x), \tag{1}$$

where $x = (x_1, x_2, \ldots, x_n)$ and $Q = \{x \in E^n \mid Ax = b, x \geq 0\}$. The function of $f(x)$ is assumed to be concave and well defined over the convex polyhedron Q. It is also assumed that the contrained minimum of $f(x)$ is finite.

The optimum solution to (1) is characterized by its occurence at an extreme point of Q. However, the principal difficulty is that a local minimum is not necessarily global.

A method for solving (1) was proposed by Hoang Tuy [12], but with the additional requirement that $f(x)$ be concave over all $x \in E^n$. Tuy's algorithm is started by identifying a local minimum point, $\bar{x}$, of Q. A hyperplane cut (called Tuy's cut) is then determined and augmented to the problem so that all feasible (extreme) points in Q having a worse value than $f(\bar{x})$ are excluded. Informally, Tuy's cut is generally defined by a hyperplane passing through the end points of the extended halflines emanating from the current local minimum such that the associated values of $f(x)$ at these end points is equal to $f(\bar{x})$. It is clear that $f(\bar{x})$ is an upper bound on the optimal objective value and that any extreme point $x \in Q$ having $f(x) \geq f(\bar{x})$ cannot be promising. The process is then continued by searching for a local minimum of the new solution space resulting from the application of the last Tuy's cut. If no new local minima exist, the algorithm is terminated with the last local minimum being the global optimum. Tuy provides no convergence proof for the algorithm.

This paper presents a new algorithm for solving (1). The algorithm is basically a branch-and-bound method which utilizes a special cutting plane procedure to identify the global extreme point of Q. The main difference between this work and Tuy's is that the cuts are generated solely from the geom-

etry of the convex polyhedron Q. Also, the identification of the candidate extreme points of Q necessitates defining a linear function which underestimates $f(x)$. The linearity restriction is important since, as will be seen later, it reduces the problem to solving a series of linear programs.

Although a method is given for developing a linear underestimator for the general case, illustrations for developing more efficient (or tighter) underestimators are also provided for important concave minimization problems.

In section II, the generalized branch-and-bound algorithm is presented and its relationship to work by other authors is discussed. Section III introduces the cutting plane method associated with the algorithm. Section IV develops a general linear underestimator for $f(x)$ and shows how "tighter" underestimators are developed for the fixed-charge problem, the separable programming problem, and the 0–1 mixed integer linear problem. Finally, section V illustrates the computational efficiency of the proposed algorithm as applied to the fixed-charge problem.

II. THE ALGORITHM

The general idea of the algorithm is explained as follows. Let $l(x)$ be a linear underestimator of $f(x)$ over Q; that is,

$$l(x) \leq f(x), \qquad x \in Q, \tag{2}$$

then it is clear that

$$\min_x \{l(x) \mid x \in Q\} \leq \min_x \{f(x) \mid x \in Q\}. \tag{2'}$$

This means that by starting with the extreme point x^0 satisfying $\min\{l(x) \mid x \in Q\}$, $\underline{f} = l(x^0)$ is a lower bound on the optimum objective value of (1), while, from (2), an obvious upper bound is given by $\bar{f} = f(x^0)$. Now consider $x^1 (\neq x^0)$, an adjacent extreme point to x^0 such that $l(x^1)$ yields the smallest $l(x)$ among all the adjacent extreme point of x^0. It is clear that only those adjacent point having $\underline{f} \leq l(x) \leq \bar{f}$ need be considered in determining x^1. The point x^1 is then said to be the next ranked extreme point. (The exact details of the general (cutting plane) procedure for determining the next ranked extreme points will be presented in section III.) Now, the new lower bound is $\underline{f} = l(x^1)$. The upper bound $\bar{f}$ is changed to $f(x^1)$ *only if* $f(x^1)$ is smaller than the *current* upper bound $\bar{f} = f(x^0)$.

In general, suppose $E^{i-1} = \{x^0, x^1, \ldots, x^{i-1}\}$ is the set of (nonredundant) extreme points thus far ranked. Then x^i, the next ranked extreme point, is selected as the adjacent extreme point to one of the elements in E^{i-1} such that $l(x^i)$ is again the smallest among all such adjacent extreme points and provided $x^i \cap E^{i-1} = \phi$, that is, x^i is nonredundant with respect to E^{i-1}. The current lower bound is now given by $\underline{f} = l(x^i)$, but the upper bound is changed to $f(x^i)$ only if this quantity is smaller than the best available upper bound $\bar{f}$.

The termination of the procedure is effected at x^k if $\underline{f} = l(x^k) \geq \bar{f}$ with the extreme point associated with $\bar{f}$ being the optimum. This follows since from (2),

$$f(x) \geq l(x^k) \geq \bar{f}, \qquad x \in Q^* - E^k,$$

where Q^* is the set of extreme points of Q. This condition shows that all the remaining extreme points $(Q^* - E^k)$ can only yield worse objective values than $\bar{f}$, and are thus nonpromising.

The above discussion can be summarized in an algorithmic form as follows:

STEP 0: Solve the linear program:

$$\min_x \{l(x) | x \epsilon Q\}$$

and let x^0 be the optimum extreme point. Define $\underline{f}(x^0) = l(x^0)$ as the lower bound on the optimum objective value of (1). Let $x^* = x^0$. Then $\bar{f}(x^*)$ is an upper bound. Set $i = 0$, then go to Step 1.

STEP 1: The current upper and lower bounds are given by $\bar{f}(x^*)$ and $\underline{f}(x^i)$, respectively. Let x^{i+1} be the next ranked extreme point of Q and set the new lower bound $\underline{f}(x^{i+1}) = l(x^{i+1})$. Go to Step 2.

STEP 2: If $l(x^{i+1}) \geq \bar{f}(x^*)$, stop; x^* is optimum. If $f(x^{i+1}) < \bar{f}(x^*)$, set $x^* = x^{i+1}$, $\bar{f}(x^*) = f(x^{i+1})$. Otherwise, the upper bound remains unchanged. Set $i = i + 1$ and go to Step 1.

The general idea of the above algorithm was first proposed by Katta Murty [7] for solving the fixed-charge problem. Murty also indicated [7, Corollary 1] that for $f(x) = D(x) + z(x)$, where $z(x)$ is linear and $D(x)$ is concave, if $l(x)$ is taken equal to $z(x)$, the algorithm is equally applicable. However, it is clear that Murty's corollary is true only if $z(x) \leq z(x) + D(x)$. This obviously is not valid, in general. Later, Cabot and Francis [3] utilized the exact algorithm to solve the case where $D(x)$ is a negative (semi)definite quadratic form. (See section IV.) The Cabot-Francis paper, however, presents the details of Murty's algorithm in a more explicit manner.

The ranking procedure of Step 1, as advanced by Murty, determines the adjacent extreme points to each element (basis) in E^i as the (new) basic solutions in which one of the current (eligible) nonbasic variables is made basic. This requires carrying out a single pivot operation as in the simplex method. The major drawbacks of Murty's procedure is that the number of generated adjacent extreme points may become very large to the extent of taxing the computer memory. Moreover, because the same (adjacent) extreme point may be generated from more than one element in E^i, a procedure is needed to avoid storing redundant points. The extensive experimentation by this author shows that Murty's algorithm, as applied to the zero–one problem, generally yields very discouraging results (See [10]). The complex bookkeeping procedures required to economize the utilization of the computer memory and to minimize redundancy shows distinctly that the algorithm can very easily reach an unmanageable state.

This paper differs from the work of Murty in two aspects:

(i) It presents a general algorithm which solves any problem of type (1). This is in contrast with Murty's (or Cabot and Francis') work which leaves the impression that it can handle specialized concave problems only.

(ii) It develops a new procedure for the details of Step 1 of the algorithm which improves on the drawback of Murty's ranking scheme. The new procedure utilizes a cutting plane technique which uses the "convexity cuts" recently developed by Glover [5].

It must be noted that, by using Murty's ranking procedure, the requirement that the underestimator $l(x)$ be linear is needed only in Step 0. This follows since the theory of linear programming automatically allows the determination of the proper extreme point, x^0. Clearly, the linearity assumption is not needed in the ranking procedure of Step 1. This is in contrast with the new ranking procedure, to be introduced in the next section, where the linearity of the underestimator is a mandatory requirement. This follows since the ranked extreme points are determined by applying the dual simplex method of linear programming.

III. A CUTTING PLANE METHOD FOR RANKING THE EXTREME POINTS OF Q

Informally, the idea of the new ranking scheme is explained as follows. Start with x^0 obtained at Step 0. Then define a cut that eliminates x^0 only from among all the extreme points of Q. The hyperplane associated with the cut is determined to pass through the adjacent extreme points of x^0. Now, augmenting the linear programming problem with the cut and applying the dual simplex method, the resulting optimum feasible solution yields the next ranked extreme point. A new cut can now be generated from the adjacent extreme points by using the new solution space. The process is repeated as necessary.

The above procedure is tailored after a recent development by Glover [5] who lays a general theory for constructing legitimate cuts which can be used systematically to determine certain points in a given convex polyhedron. A typical illustration is the convex polyhedron Q with its extreme points representing the set of points to be identified. Glover's theory actually generalizes earlier ideas by Young [13] and Balas [1] where they developed legitimate cuts for the integer linear programming problem.

To formalize the above discussion, let the current basic solution be defined by the set of equations

$$y_i = b_{i0} - \sum_{j \in N} b_{ij} t_j, \qquad i \in M,\ y_i,\ t_j \geq 0, \tag{3}$$

where y_i and t_j are the basic and nonbasic variables, respectively. The sets M and N define the indices of the basic and nonbasic variables. The cut referred to above is now described based on Glover's theory:

Glover's Convexity Cut Lemma [5]: Let S be a set of points in the convex polyhedron Q. If R is a convex set whose interior contains no point in S, and if $y_i = b_{i0}$, $i \in M$ (possibly a boundary point of R) has a deleted feasible neighborhood which lies in the interior of R, then for any constant $t_j^* > 0$, $j \in N$, such that

$$y_i = b_{i0} - b_{ij} t_j^* \in R, \qquad \forall\ i \in M,$$

the convexity cut †

$$\sum_{j \in N} t_j / t_j^* \geq 1 \tag{4}$$

excludes the extreme point $y_i = b_{i0}$, $i \in M$, but never any point in S.

The application of the above lemma to Step 1 of the algorithm is straightforward. Here the set S consists of the unranked extreme points of the current solution space. The point $y_i = b_{i0}$, $i \in M$, takes the place of the current "ranked" extreme point, and the set R is represented by the convex polyhedron describing the current solution space.

The determination of the constants t_j^* in (4) follows directly from the theory of the simplex method; that is,

† Because of the convexity requirement stipulated on the Set R, Glover coins the suggestive name "convexity cut."

$$t_j^* = \begin{cases} \min_{i \in M} \dfrac{b_{i0}}{b_{ij}}, & b_{ij} > 0 \\ \infty, & \text{all } b_{ij} \leq 0. \end{cases}$$

Clearly, t_j^* is strictly positive if the current solution is nondegenerate, that is, $b_{i0} > 0$. When $b_{i0} = 0$ for at least one $i \in M$, then it is possible that $t_j^* = 0$ and the convexity cut (4) becomes undefined.

In order to overcome the above difficulty resulting from a degenerate situation, we use the following procedure due to Balas [1].†

Degeneracy occurs when an extreme point is "overdetermined," that is, when the current solution point has more than n hyperplanes associated with it, where n is the total number of variables. Balas [1] proves that by dropping each constraint for which the associated basic variable is equal to zero, the resulting convex polytope necessarily associates n distinct edges with the current solution vertex. Under this condition, the values of t_j^* are readily determined. Of course, when the cut is added, all the deleted constraints must be activated before the problem is reoptimized; unless such constraints are proved to be redundant with respect to R in which case they can be dropped completely.

There are two important points which must be considered in association with the degeneracy problem. These difficulties do not arise in Balas' case mainly because the sets R and S in his problem remain unaffected by the deletion of the constraints associated with the zero basic variables. This obviously is not the case in our situation.

(i) Let C be the degenerate cone associated with the current solution vertex, $\bar{X}$, and define L as the polytope obtained from the current solution space of the problem by deleting the halfspaces associated with C. Further, define C' as the nondegenerate cone associated with $\bar{X}$ which is obtained from C by deleting all the constraints satisfying Balas' condition.

Since $C \subset C'$, then the cut obtained from the adjacent extreme points resulting from the intersection of C' with L cannot be stronger than its equivalence when C replaces C'. This means that the new cut cannot eliminate any of the extreme points of $R (= C \cap L)$ which have not been tested for optimality. Subsequently, the cut obtained by using C' is legitimate with respect to R.

(ii) The cut obtained by using C' will most likely create new extreme points which are not part of the vertices of the original solution space. The question then arises as to the possibility of the optimal solution being "trapped" at one of these vertices. This point is refuted as follows:

By the convexity cut lemma, such extreme points (when they occur) must lie on the halfline

$$y_i = b_{i0} - b_{ij} t_j, \qquad 0 < t_j < t_j^*, \qquad i \in M. \tag{5}$$

If (5) is an edge (or segment thereof) of the original solution space Q (defined in (1)), then the new extreme point is actually a nonvertex of Q. Consequently, it cannot yield an improved solution point as this leads to contradiction. A similar argument applies if (5) is a new edge resulting from the application of previous cuts.

---

†It must be noted that Murty's procedure overcomes the degeneracy problem by enumerating all the basic solutions associated with the current extreme point. From the computational point of view, this has proved to be very time consuming (see [10]).

It is important to notice that the effect of degeneracy goes beyond simple inconvenience in computation. Essentially, the creation of new extreme points must reduce the efficiency of the proposed method since it may be necessary to test these points for optimality (see the numerical example in section VI for an illustration). Consequently, it seems important that serious consideration must be given to minimizing the effect of degeneracy. The work of Thompson, Tonge, and Zionts [11] provides ways for eliminating degeneracy in certain situations (as illustrated by the numerical example in section IV). However, there does not yet exist a general method for handling all degeneracy situations.

IV. DETERMINATION OF THE LINEAR UNDERESTIMATOR $l(x)$

In this section we show how a linear underestimator $l(x)$ can be developed for $f(x)$ in the general case. Since the efficiency of the proposed algorithm should depend on the selection of the underestimator, illustrations showing how tighter underestimators can be developed for an important class of concave minimization problems. This includes the fixed-charge problem, the separable programming problem, the quadratic problem, and the 0–1 mixed integer problem.

(i) General Underestimator $l(x)$:

From the properties of the concave functions, a tangent hyperplane to $f(x)$ at $\bar{x}$ [assume $\bar{x}\epsilon Q$, where Q is the convex polyhedron defined in (1)] overestimates $f(x)$. Consequently, it appears plausible that we can make use of a tangent hyperplane to $g(x)=-f(x)$ (with modifications) to underestimate $f(x)$. Let $t_g(x)$ be a tangent hyperplane to $g(x)$ at a given point.† Clearly, for any x,

$$t_g(x) \leq g(x). \tag{6}$$

Now a transition from $g(x)$ to $f(x)$ can be made if $g(x) \leq f(x)$. Unfortunately, this is not true in general. However, if the values of x are restricted to those in Q, then the transition can be achieved as follows:

PROPOSITION: Let $M \geq 0$ be a real number, then there must exist a value of $M < \infty$ such that

$$-M + t_g(x) \leq f(x), \qquad x\epsilon Q. \tag{7}$$

In this case $l(x) = -M + t_g(x)$.

PROOF: We need only show that $-M+g(x) \leq f(x)$ for all $x\epsilon Q$. The minimum (maximum) value of $f(x)\,(g(x))$ occurs at an extreme point of Q. If $\min_{x\epsilon Q} f(x) \geq 0$ then $\max_{x\epsilon Q} g(x) \leq 0$ and obviously the desired result is achieved for $M=0$. Now, suppose $\min_{x\epsilon Q} f(x) < 0$, then $\max_{x\epsilon Q} g(x) > 0$. By assumption, $f(x)$ possesses a finite minimum over Q. Thus M can be selected such that $M \geq |\min_{x\epsilon Q} f(x)|$. Since, by symmetry, $|\min_{x\epsilon Q} f(x)| = \max_{x\epsilon Q} g(x)$ it follows that,

$$-M+g(x) \leq -M+\max_{x\epsilon Q} g(x) \leq 0 \leq M+\min_{x\epsilon Q} f(x) \leq M+f(x), \qquad x\epsilon Q.$$

Now, since $M \geq |\min_{x\epsilon Q} f(x)|$ can be taken arbitrarily large, then letting $M \geq 2|\min_{x\epsilon Q} f(x)|$, the desired conclusion follows immediately.

† We further assume that the tangent hyperplane is determined at $\bar{x}$ satisfying $\nabla g(\bar{x}) \neq 0$, where $\nabla g(\bar{x})$ is the gradient vector of $g(x)$ at $\bar{x}$. This will ensure that the resulting linear underestimator is not trivial.

The above proposition actually implies that a linear underestimator for $f(x)$, $x \epsilon Q$, can be taken as

$$l(x) = -\sum_{j=1}^{n} m_j x_j - M,$$

where m_j are positive constants and $M \geq |\min_{x \epsilon Q} f(x)|$. If $\min_{x \epsilon Q} f(x) \geq 0$, then M can be taken equal to zero.

Notice that since the lower bound on M is obviously not known a priori, one must rely on some practical estimate to determine a numerical value for M. Although any values of $m_j > 0$ can be utilized, further research is needed to determine the best set of values providing the tightest linear underestimator.

(ii) Fixed-Charge Problem:

In the fixed-charge problem, $f(x)$ is defined by

$$f(x) = \sum_{j \epsilon N} c_j x_j + \sum_{j \epsilon N} K_j \delta(x_j), \qquad N = \{1, \ldots, n\},$$

where $\delta(x_j) = 0$ if $x_j = 0$, and $\delta(x_j) = 1$ if $x_j > 0$. The coefficients c_j and K_j are real numbers with $K_j > 0$ for all j. It can be proved that $f(x)$ is a concave function which is continuous everywhere except at $x = 0$. Now, since $K_j > 0$ by assumption, it follows that

$$\sum_{j \epsilon N} c_j x_j \leq \sum_{j \epsilon N} c_j x_j + \sum_{j \epsilon N} K_j \delta(x_j), \qquad x_j \geq 0.$$

This shows that the linear underestimator can be taken as

$$l(x) = \sum_{j \epsilon N} c_j x_j. \tag{8}$$

Notice $l(x)$ is valid for any $x_j \geq 0$.

The application of the above estimator will be illustrated numerically in the next section. Notice that the same idea can be utilized to solve certain problems that often arise in inventory theory. A typical example is the finite horizon, multiple-item model in which price breaks (or quantity discounts) are allowed in the ordering function. This typically results in a piecewise-linear concave function. In this case, the linear segments representing the smallest per unit ordering cost can be used to determine $l(x)$.

(iii) Separable Programming Problem:

Let

$$f(x) = \sum_{j=1}^{n} f_j(x_j)$$

where $f_j(x_j)$ is a concave and well defined function. Suppose now that the feasible range for each x_j as defined by the solution space Q is given by $a_j \leq x_j \leq b_j$, where a_j and b_j are known constants. Let $l_j(x) = \alpha_j x_j + \beta_j$ be the straight line joining the two points $(a_j, f_j(a_j))$ and $(b_j, f_j(b_j))$. Since $f_j(x_j)$ is concave, it follows by definition that

$$l_j(x_j) \leq f_j(x_j), \qquad x_j \epsilon Q.$$

Consequently,

$$l(x) = \sum_{j=1}^{n} l_j(x_j). \tag{9}$$

It is noted that the fixed-charge problem discussed in (ii) satisfies the condition for a separable concave problem. Consequently, the above linear underestimator can also be used with the fixed-charge problem. Notice, however, that the present underestimator is tighter than that defined in (ii). This follows since it is defined for constrained values of x only as compared with $x \geq 0$ in the fixed-charge problem.

(iv) Quadratic Minimization Problem:

Let

$$f(x) = z(x) + D(x),$$

where $z(x)$ is linear and $D(x)$ is a negative (semi)definite quadratic form. If $D(x) = xBx^T$, then B is negative (semi)definite. The linear underestimator in this case was developed by Gilmore [4] and by Lawler [6] and was subsequently utilized by Cabot and Francis [3] in connection with Murty's algorithm.

Let b_j be the jth column of B. Then $D(x)$ can be written as

$$D(x) = \sum_{j=1}^{n} (xb_j)x_j.$$

Define $u_j = \min_{x \in Q} \{xb_j\}$. Hence

$$l(x) = z(x) + \sum_{j=1}^{n} u_j x_j \leq z(x) + D(x). \tag{10}$$

Notice that since $x_j \geq 0$, for all j, then, for any finite $w_j \leq u_j$, $z(x) + \sum_{j=1}^{n} w_j x_j$ still provides a legitimate underestimator. This result can sometimes be used advantageously to avoid solving n linear programs. An application of this situation is given below.

(v) Zero-One Mixed Integer Problem:

In this problem

$$f(x) = \sum_{j \in N} c_j x_j, \qquad x_j = (0, 1) \qquad \text{for } j \in N^1 \subset N.$$

The function $f(x)$ can be written equivalently as,

$$f(x) = \sum_{j \in N - N^1} c_j x_j + \sum_{j \in N^1} (c_j + M(1 - x_j))x_j, \qquad M > 0 \text{ and very large.}$$

$$= \sum_{j \in N} c_j x_j + M \sum_{j \in N^1} x_j + M \sum_{j \in N^1} (-x_j^2),$$

where $x_j \geq 0$, $j \epsilon N$, and $x_j \leq 1$, $j \epsilon N^1$. The expression $M(1-x_j)$ assigns a very high penalty to x_j for $0 < x_j < 1$, $j \epsilon N^1$, thus allowing it to take binary values only. The mixed integer objective function has been equivalently converted into a quadratic function in which the quadratic form $M \sum_{j \epsilon N^1} (-x_j^2)$ is clearly negative definite. Notice that all the variables in the new form are continuous.

The above equivalence relationship was developed by Raghavachari [8] and independently by Taha [10] in an effort to secure a simpler formulation for the mixed 0–1 integer problem.

The transformed $f(x)$ is exactly in the same form as the function in (iv). Thus, using the method in (iv), u_j, $j \epsilon N^1$, is defined by

$$u_j = \min_{x_j} \{ -Mx_j | x_j \epsilon Q, \quad 0 \leq x_j \leq 1 \} \tag{11}$$

$$\geq \min_{x_j} \{ -Mx_j | 0 \leq x_j \leq 1 \} = -M.$$

Thus, taking $u_j = -M$, it follows from the development in (iv) that $l(x) = \sum_{j \epsilon N} c_j x_j$. This shows that $l(x)$ can be taken as $f(x)$ after removing the condition $x_j = (0, 1)$, $j \epsilon N^1$. Notice that if u_j is determined from the exact linear program in (11), the main difference would be that $0 \leq x_j < 1$ for some $j \epsilon N^1$, that is, $x_j = 0$. This means that the new $l(x)$ will be the same as above except that the indicated x_j are set equal to zero.

The above result can also be derived on intuitive basis. Since for the 0–1 mixed integer problem the optimum must occur at an extreme point, the integrality condition can be replaced by the continuous range $0 \leq x_j \leq 1$. It then follows that $\left\{ \sum_{j \epsilon N} c_j x_j | 0 \leq x_j \leq 1, j \epsilon N^1 \right\}$ must underestimate $\left\{ \sum_{j \epsilon N} c_j x_j | x_j = (0, 1), j \epsilon N^1 \right\}$ since the former is less restrictive. This, incidentally, means that the transformation of $f(x)$ given above does not yield any privileged information and hence is trivial.

Notice that by using the continuous range $0 \leq x_j \leq 1$, $j \epsilon N^1$, the resulting objective function becomes linear in x_j over its feasible values. Thus the new objective function may be considered concave over the feasible space and the general algorithm in section II becomes applicable. In this case the upper bound $\bar{f}$ is defined equal to ∞ for any extreme point not satisfying $x_j = (0, 1)$, $j \epsilon N^1$. The important point, however, is that the cut as defined in section III is uniformly weaker than its equivalence as developed by Balas [1]. On the other hand, the determination and use of Balas' stronger cut requires more complex computation as compared with ours. Consequently, the real merit of either cut can only be checked through computational experimentation.

V. COMPUTATIONAL EXPERIENCE WITH THE FIXED-CHARGE PROBLEM

This section illustrates the efficiency of the proposed algorithm by applying it to the fixed-charge problem. This special case is selected primarily because of its practical interest. In addition, the availability in the literature of computational results for other fixed-charge methods allows a more meaningful evaluation of the proposed algorithm.

In order to clarify the details of the algorithm, especially those associated with the degeneracy problem, we first introduce a numerical example. This will be followed by a presentation of the computer results as applied to randomly generated problems.

EXAMPLE:

$$\text{minimize } f(x) = \phi_1(x_1) + \phi_2(x_2)$$

subject to

$$2x_1 + x_2 + S_1 = 4$$

$$x_1 + x_2 + S_2 = 3$$

$$0 \leq x_2 \leq 5/2$$

$$x_1, S_1, S_2, \leq 0,$$

where

$$\phi_1(x_1) = \begin{cases} 0, & x_1 = 0 \\ -4x_1 + 1, & x_1 > 0 \end{cases}$$

$$\phi_2(x_2) = \begin{cases} 0, & x_2 = 0 \\ -3x_2 + 1/2, & x_2 > 0. \end{cases}$$

Thus, $l(x) = -4x_1 - 3x_2$. Using Dantzig's technique to accommodate the upper bound on x_2, Table I gives the solution specifying x^0. A graphical display of the solution is given in Figure 1.

TABLE I

		S_1	S_2	
$l =$	-10	-1	-2	$x^0 = (1, 2)$; Point ⓪
				$f(x^0) = -10$
$x_1 =$	1	1	-1	$\bar{f}(x^0) = -10 + (1 + 1/2) = -8\,1/2$
$x_2 =$	2	-1	2	$\bar{f}(x^*) = \bar{f}(x^0) = -8\,1/2$

Cut #1 is now developed. (Notice that the determination of the constants (t_j^*) of the cut must be based on Dantzig's upper bounding technique.) Thus,

$$S_1^* = \min\left\{1/1, \frac{2 - 5/2}{-1}, \infty\right\} = 1/2$$

$$S_2^* = \min\{2/2, \infty, \infty\} = 1$$

and the cut is given by

$$\frac{S_1}{1/2} + \frac{S_2}{1} \geq 1.$$

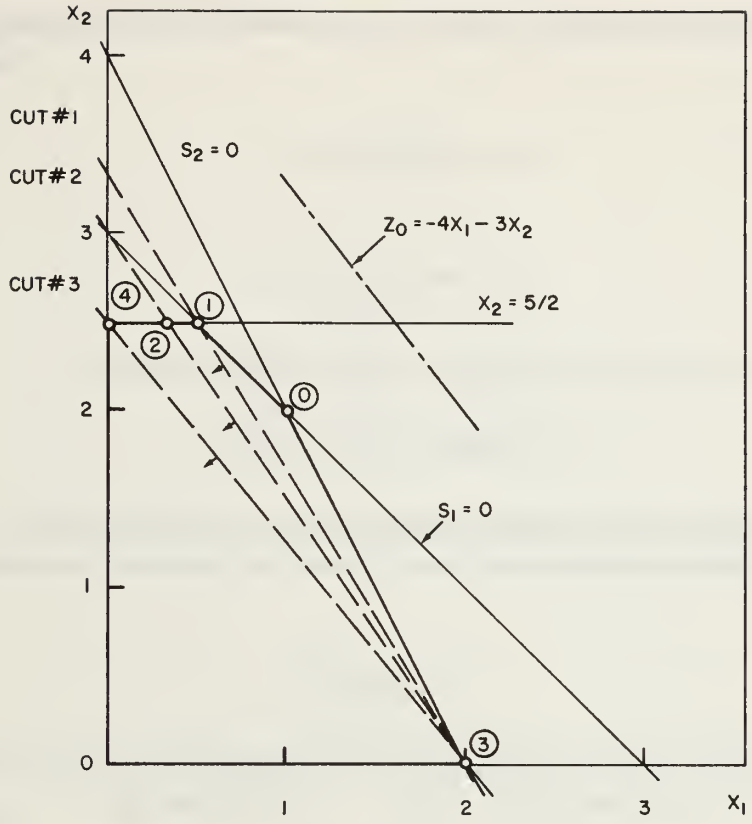

FIGURE 1. Solution of the numerical example.

Expressed in terms of x_1 and x_2, the cut is

(Cut #1) $$5x_1 + 3x_2 \leq 10;$$

We denote the slack as S_3.

Table II yields x^1 as a result of augmenting Table I by Cut #1 and reoptimizing using the dual simplex method for upper bounded variables.

TABLE II

		S_2	S_3	
$l=$	$-9\frac{1}{2}$	$-3/2$	$-1/2$	$x^1=(\frac{1}{2}, \frac{5}{2})$; Point (1)
$x_1=$	$1/2$	$-3/2$	$1/2$	$\underline{f}(x^1)=-9\frac{1}{2}$
$x_2=$	$5/2$	$5/2$	$-1/2$	$\bar{f}(x^1)=-9\frac{1}{2}+(1\frac{1}{2})=-8>\bar{f}(x^*)$
$S_1=$	$1/2$	$1/2$	$-1/2$	$\bar{f}(x^*)=f(x^0)$

Notice that in Table II, x_2 is basic at its upper bound. This means that the current solution is degenerate. Using Balas' condition which, in this case, calls for ignoring the equations involving basic vari-

ables at upper bound or zero level, it is clear that the x_2-equation must be disregarded in developing cut #2. Thus,

$$S_2^* = \min\left\{\frac{1/2}{1/2}, \infty, \infty\right\} = 1$$

$$S_3^* = \min\left\{\frac{1/2}{1/2}, \infty, \infty\right\} = 1$$

This yields a new cut which when expressed in terms of x_1 and x_2 is given by

(Cut #2) $$6x_1 + 4x_2 + S_4 = 12, \qquad S_4 \geq 0.$$

Table III gives the new solution after Cut #2 is effected. Notice that $x_2 = 5/2 - x_2'$. Notice also that since S_3 is associated with a previous cut and since it is basic, its corresponding equation can be dropped in future tableaus.

TABLE III

		x_2'	S_4	
$l =$	$-8\frac{5}{6}$	$-1/3$	$-2/3$	$x^2 = (\frac{1}{3}, \frac{5}{2})$; Point ②
$x_1 =$	1/3	$-2/3$	1/6	$\underline{f}(x^2) = -8\frac{5}{6}$
$S_2 =$	1/6	$-1/3$	$-1/6$	$\bar{f}(x^2) = -8\frac{5}{6} + 1\frac{1}{2} = -7\frac{1}{3} > \bar{f}(x^*)$
$S_1 =$	5/6	1/3	$-1/3$	$\bar{f}(x^*) = \bar{f}(x^0)$
$S_3 =$	5/6	1/3	$-5/6$	

Cut #3 is now generated from Table III. This gives

(Cut #3) $$30x_1 + 24x_2 \leq 60.$$

The application of this cut will yield point ③ with $x^3 = (2, 0)$ and $\underline{f}(x^3) = -8$. Since $\underline{f}(x^3) > \underline{f}(x^*)$, the process terminates. Thus $x^* = x^0$ is the optimum solution.

Notice the effect of degeneracy at ①. Point ① is (over) determined by the three lines $x_2 = 5/2$, $x_1 + x_2 = 3$, and $6x_1 + 4x_2 = 12$. Balas' condition drops $x_2 = 5/2$. The cone C' as introduced in section III, is then defined by the halfplanes $x_1 + x_2 \leq 3$ and $6x_1 + 4x_2 \leq 12$, which yields Cut #2. The optimum point ② is a new extreme point which does not belong to the original solution space.

It is remarked that if the redundant constraint $x_1 + x_2 \leq 3$ is eliminated instead of $x_2 \leq 5/2$, then Cut #2 would have been stronger as it would pass through points ③ and ④. Stanley Zionts, in a private communication to the author, shows that by using the results in [11], this specific degeneracy situation can be avoided. The idea is as follows: Prior to constructing a cut constraint, if there is any degeneracy write the degenerate constraint so that the right-hand side element is zero. (In [11], methods for identifying redundant, (and of course redundant degenerate) constraints are provided.) Applying this to Table II, x_2 is replaced by $5/2 - x'$. In order for the redundant constraint to be implied in "defi-

nitional" form, x_2' must now be made nonbasic with S_2 being the new basic variable. This yields Table IV (Table II revised).

TABLE IV

		x_2'	S_3
l	$-9\frac{1}{2}$	$-3/5$	$-4/5$
x_1	$\frac{1}{2}$	$-3/5$	$1/5$
S_2	0	$-2/5$	$-1/5$
S_1	$\frac{1}{2}$	$1/5$	$-2/5$

Notice that the S_2-row is redundant now and may be dropped from the tableau. But more importantly, the generated cut is

$$\frac{x_2'}{5/2}+\frac{S_3}{5/2}\geq 1,$$

which now passes through points ③ and ④, thus bypassing the extra point ② and its associated cut.

Computer Results

The testing of the algorithm as applied to the fixed charge problem is designed to check the effect of the size of the problem and the magnitude of the fixed charge on the speed of computation. Random problems of the type

$$\max\{\sum_{j=1}^{n}(c_jx_j+K_j\delta(x_j))\,|\,\sum_{j=1}^{n}a_{ij}x_j\leq b_i,\ x_j\geq 0,\ i=1,\ \ldots\ m\}$$

are generated with their coefficients lying in the ranges

$$\begin{aligned}0\leq c_j &\leq 999\\ 0\leq K_j &\leq 160\\ -20\leq a_{ij} &\leq 100\\ 0\leq b_i &\leq 200.\end{aligned}$$

The sizes of the generated problem are given by $(m\times n)=(5\times 20)$, (5×30), (10×20), and (15×30). In order to test the effect of the fixed charge, the same problems are used again with K_j replaced by $2K_j$ and $3K_j$, respectively. No special structure is specified for the problems and the density of the matrix $A=\|a_{ij}\|$ is at least 97 percent.

The algorithm is coded in FORTRAN IV for the IBM 360/50. The results are summarized in Table I.

One of the basic difficulties we encountered in coding the algorithm was the control of machine round-off error. This is important since a zero variable may be rounded to a positive value, thus affecting

TABLE V. *Summary of Computation*

(Time in seconds)

Problem number	$(m \times n) = (5 \times 20)$			$(m \times n) = (5 \times 30)$			$(m \times n) = (10 \times 20)$			$(m \times n) = (15 \times 30)$		
	K_j	$2K_j$	$3K_j$	K_j	$2K_j$	$3K_j$	K_j	$2K_j$	$3K_j$	K_j	$2K_j$	$3K_j$
1	0.300	0.500	10.70	0.484	0.513	0.683	64.917	65.117	150.317	40.600	44.150	702.500
2	0.216	0.250	0.250	0.366	0.467	23.650	22.833	22.950	127.950	9.184	11.400	11.167
3	0.250	0.283	0.450	0.467	0.483	0.483	1.417	192.117	192.117	59.000	66.183	72.350
4	0.216	0.717	1.717	1.530	2.633	3.116	70.650	77.117	81.980	3.650	417.183	702.500
5	0.317	0.317	2.783	0.866	0.750	1.550	1.550	2.000	33.967	19.984	135.433	803.117
Average	0.260	0.413	3.180	0.742	0.969	5.896	32.273	71.86	97.266	26.482	134.868	327.552

the bounds directly. The problem was overcome by using double-precision computation as well as appropriate tolerances. Also, checks were implemented in the code to detect the accumulation of machine round-off error. For example, an important check is to test whether at a given iteration the number of positive variables among the *original* variables exceeds the number of *original* constrainst. It must be remarked that the five problems in Table V were selected from among 20 test problems as the ones yielding the least amount of "disorder" from the viewpoint of machine round-off error. The remaining problems were excluded by the checks in the code because they indicated uncontrollable round-off error. It is felt, however, that a professional programmer should be able to develop a more efficient and accurate code than the one written by the author.

Although the results in Table I are generally compatible with what one may expect; that is, the average computation time increases with the increase in the fixed charges, the individual problems are exhibiting peculiar behavior which needs explanation. For example, problem $3-(10 \times 20)$ requires 1.5 seconds for K_j, 192 seconds for $2K_j$, and again 192 seconds for $3K_j$. This result can be justified as follows:

The termination of the algorithm occurs at the extreme point x^i when $l(x^i) \geq \bar{f}(x^*)$. It is obvious that the computation time of the problem is primarily a function of the number of extreme points which are ranked before termination occurs. Thus, two problems *having the same solution space*, will require the same computation time if they terminate at the same x^i. Notice that $l(x)$ is dependent only on the linear terms of the objective function and that its value at an extreme point is not dependent on the fixed charges, while $\bar{f}(x^*)$ is directly dependent on the fixed charges. Consequently, if $l(x^i) - \bar{f}(x^*)$ for $2K_j$ is large enough to accommodate an increase in the fixed charges to $3K_j$, termination still occurs at x^i and the same computation time is consumed. Similarly, if $l(x^i) - \bar{f}(x^*)$ for K_j is too small, an increase in the fixed charges to $2K_j$ may necessitate further ranking of new extreme points before termination is effected.

The results in Table I also show that the computation time increases more appreciably with the increase in the number of constraints rather than with the number of variables. These results differ from those associated with cutting plane algorithms in integer programming where the number or variables is the main factor affecting the computation time. The reason for this appears to be that our

algorithm depends more directly on the number of extreme points of the solution space which is a function of both the number of constraints and the number of variables.

For the sake of comparing our algorithm with other exact methods for the fixed charge problem, we only came across two algorithms by Bod [2] and Steinberg [9]. The two methods are of the branch and bound type. Bod's method utilizes what may be termed as a partial enumeration technique for testing all the extreme points (basic feasible solutions) of the convex polyhedron. The effective use of bounds on the objective value excludes most of the nonpromising extreme points. Steinberg's method, on the other hand, initiates two problems at each node according to whether the variable x_j is zero or positive. Bounds on the objective value are also used to effect the proper termination of the algorithm.

Bod does not present computer results for his algorithm. But Steinberg tests two sets of problem with sizes (5×10) and (15×30) on the IBM 360/50. The average computation times per problem for the two sets are 10 sec and 21.1 min, respectively. This is far inferior to the average computation time obtained by our algorithm; especially that Steinberg's algorithm can easily tax the computer memory. He reports that a set of 15 problems, with size (5×10) each, requires an average of 32 nodes while those with size (15×30) each require an average of 1,208 nodes. This shows that the number of nodes can become very large even for problems, with modest sizes. The problem is not present in our algorithm since, as in any cutting plane algorithm, the size of the matrix A at any iteration cannot exceed $(m+n)xn$.

We must remark also that, contrary to our algorithm, Steinberg's algorithm becomes slower as the magnitude of the fixed charge decreases. He utilizes the ranges $0 \leq c_j \leq 20$ and $0 \leq K_j \leq 999$ for his test problems, but does not study the effect of variations in K_j on the speed of computation.

VI. CONCLUSIONS

The algorithm presented in this paper is general in the sense that it can handle any concave minimization problem over a convex polyhedron. If the computer results of the algorithm as applied to the fixed charge problem are at all indicative of its efficiency, it would appear that the algorithm can actually be used to solve practical problems. Further research is still needed, however, to develop the tightest linear underestimator for $f(x)$. Also, since degeneracy is a pronounced problem in our algorithm, a general method is needed for treating the degenerate case without weakening the resulting cuts. This should improve the efficiency of computation considerably.

VII. ACKNOWLEDGMENT

The author wishes to thank Professor Stanley Zionts, State University of New York at Buffalo, for his helpful comments.

REFERENCES

[1] Balas, E., "Intersection Cuts—A New Type of Cutting Plane for Integer Programming," Operations Research *19*, 19–39 (1971).

[2] Bod, P., "Solution of a Fixed Charge Linear Programming Problem," *Proceedings of Princeton Symposium on Mathematical Programming* (Princeton University Press, Princeton, New Jersey, 1970), pp. 367–375.

[3] Cabot, A. V. and R. L. Francis, "Solving Certain Nonconvex Quadratic Minimization Problems by Ranking the Extreme Points," Operations Research *18*, 82–86 (1970).

[4] Gilmore, P. C., "Optimal and Suboptimal Algorithms for the Quadratic Assignment Problem," SIAM Journal *10*, 305–313 (1962).

[5] Glover, F., "Convexity Cuts and Cut Search," Operations Research, *21*, 123–134 (1973).

[6] Lawler, E. L., "The Quadratic Assignment Problem," Management Science *9*, 586–599 (1963).

[7] Murty, K. G., "Solving the Fixed Charge Problem by Ranking the Extreme Points," Operations Research *16*, 268–279 (1968).

[8] Raghavachari, M., "On the Zero-One Integer Programming Problem," Operations Research *17*, 680–684 (1969).

[9] Steinberg, D. I., "The Fixed Charge Problem," Nav. Res. Log. Quart. *17*, 217–235 (1970).

[10] Taha, H., "On the Solution of Zero-One Linear Programs by Ranking the Extreme Points," *Technical Rept. No. 71–2*, University of Arkansas (Feb. 1971) revised May 1972.

[11] Thompson, G. L., F. Tange, and S. Zionts, "Techniques for Removing Nonbinding Constraints and Extraneous Variables from Linear Programming Problems," Management Science *12*, 588–608 (1966).

[12] Tuy, H., "Concave Programming Under Linear Constraints," Soviet Math *5*, 1437–1440 (1964).

[13] Young, R. D., "New Cuts for a Special Class of 0–1 Integer Programs," Research Report, Rice University, Texas (Nov. 1968).

ESTIMATION OF A HIDDEN SERVICE DISTRIBUTION OF AN M/G/∞ SYSTEM*

Laurence Lee George

University of Louisville
Louisville, Kentucky

and

Avinash C. Agrawal

University of British Columbia
Vancouver, B.C., Canada

ABSTRACT

The maximum likelihood estimator of the service distribution function of an $M/G/\infty$ service system is obtained based on output time observations. This estimator is useful when observation of the service time of each customer could introduce bias or may be impossible. The maximum likelihood estimator is compared to the estimator proposed by Mark Brown, [2]. Relative to each other, Brown's estimator is useful in light traffic while the maximum likelihood estimator is applicable in heavy traffic. Both estimators are compared to the empirical distribution function based on a sample of service times and are found to have drawbacks although each estimator may have applications in special circumstances.

1. INTRODUCTION

Suppose customers arrive at a service system at instants $T_1, T_2, \ldots T_n$, where $\{T_n\}$ is a stationary Poisson process with rate parameter λ customers per unit time. Each customer is served upon arrival and there are sufficient servers. Service times are independently and identically distributed with some unknown distribution function $G(t)$, $t \geq 0$. These conditions describe the $M/G/\infty$ service system. They are often found in self service systems. In design of such systems it may be necessary to determine the unknown service distribution. Direct observations on the service time for each customer that enters the system may not be possible because of the economic constraints or because of other factors such as introduction of unavoidable bias, or simply, the actual behaviour of the customers while in the system is unobservable. An example of the first case may be cars entering a freeway where the distribution function of the time spent by cars on the freeway is to be estimated and tracing each car individually to find the time spent on the freeway may be extremely expensive. A similar situation may also exist in any store where it may not be possible to follow each customer through the store. Another effect of making direct observations on service time is to bias observations as customers may become conscious of being observed. It is for these reasons that direct observations on service time may not be possible. The service distribution, therefore, is hidden and estimation must be based on information other than a sample of service times.

*This research was supported in part by the Defence Research Board of Canada Grant Number 9701–25, when the authors were at the University of British Columbia.

2. MAXIMUM LIKELIHOOD ESTIMATOR OF THE HIDDEN SERVICE DISTRIBUTION WHEN λ IS KNOWN AND OBSERVATIONS ON OUTPUT TIMES ARE AVAILABLE

Mirasol [5] shows that the output of an $M/G/\infty$ service system is a nonstationary Poisson process with

(2.1) Pr (number of departures in $(0, t) = n \mid$ system initially empty)

$$= \frac{e^{-\lambda \int_0^t G(x)dx} \cdot (\lambda \int_0^t G(x)dx)^n}{n!}, \qquad n = 0, 1, \ldots,$$

where $G(\cdot)$ is the common service time distribution function and λ is the Poisson arrival rate. The intensity function of this time dependent process, $\lambda \cdot G(t)$, is both nonnegative and nondecreasing and is bounded above by λ the Poisson arrival rate.

The likelihood function for a nonstationary Poisson process with $t_1, t_2, \ldots, t_n$ as the times of occurrence of events is given by the joint density function

(2.2) $$f_{T_1, T_n, \ldots, T_n}(t_1, t_2, \ldots, t_n ; \lambda(t)) = Pr \text{ (observing events at } t_1, t_2, \ldots t_n ; \lambda(t))$$

$$= L(t_1, t_2, \ldots t_n; \lambda(t))$$

$$= \left[\prod_{k=1}^{n} \lambda(t_k)\right] \cdot \exp\left(-\int_0^{t_n} \lambda(x)\, dx\right),$$

where $\lambda(t)$ is the intensity function of the Poisson events. The first step in the problem under study involves finding a function $\lambda(t)$, $t \geq 0$ which maximizes the likelihood function given by Equation (2.1) for fixed $t_1, t_2, \ldots, t_n$ under the condition that $\lambda(t)$, $t \geq 0$, is nonnegative and nondecreasing. The maximum likelihood estimate of $\lambda(t)$, $t \geq 0$ satisfying these conditions has been obtained by Boswell [1] as,

(2.3) $$\hat{\lambda}(t) = \begin{cases} = 0 & \text{if } 0 \leq t < t_1 \\ = \min\{M, \hat{\lambda}(t_k)\} & \text{if } t_k \leq t < t_{k+1},\ k = 1, 2, \ldots, (n-1) \\ = M < \infty & \text{if } t \geq t_n, \end{cases}$$

where

(2.4) $$\hat{\lambda}(t_k) = \max_{1 \leq \alpha \leq k} \min_{k \leq \beta \leq n} \left\{ \frac{\beta - \alpha + 1}{(a_\alpha + \ldots + a_\beta)} \right\}$$

and

$$a_k = t_{k+1} - t_k, \quad k = \alpha, \alpha + 1, \ldots, \beta.$$

It may be noted that in the absence of an upper bound M on the value of $\lambda(t)$, the solution obtained will carry no meaning as (2.2) can be made arbitrarily large by setting $\lambda(t) = \epsilon > 0$ for $t < t_n$ and setting $\lambda(t_n)$ arbitrarily large. Therefore, let $\lambda(t) \leq M$ for some fixed positive number.

The maximum likelihood estimate $\hat{\lambda}(t)$ for the function $\lambda(t)$, $t \geq 0$, may be used to estimate $\lambda \cdot G(t)$, $t \geq 0$, from output observations of an $M/G/\infty$ system during some interval $[0, T]$, $T > t$. This will give an estimate of $\lambda \cdot G(t)$ for $t \in [0, T]$. To obtain estimates for $\lambda \cdot G(t)$ for small t, the output process should be observed for small t. For large values of t, the output becomes a stationary Poisson process at rate λ. For large values of t, $G(t)$ is estimated as 1. If the input rate λ is assumed to be known (this may be estimated from input data) and it is also assumed that the system starts empty, then the maximum likelihood estimate of the service distribution function $G(t)$ is given by:

$$\hat{G}(t) = \min\left[\frac{\hat{\lambda}(t)}{\lambda}, 1\right]. \tag{2.5}$$

In case it is desired to relieve the assumption that the system starts emtpy at $t=0$, one must consider the first outputs as order statistics from $G(t)$, given the number in the system at $t=0$ possibly mixed with outputs which arrive after $t=0$.

The hidden service distribution $G(t)$ for an $M/G/\infty$ system may also be estimated by peeking at the system only at times $t_1, t_2, \ldots t_n$ and observing the number in the system $N(t_1), \ldots, N(t_n)$. This sequence may also be used for maximum likelihood estimation because the number of customers in the $M/G/\infty$ service system is also a nonstationary Poisson process with intensity function $\lambda(1-G(t))$ nonincreasing in t. The maximum likelihood estimator $\hat{\lambda}(t)$ from (2.4) may be made into a maximum likelihood estimator of a nonincreasing function by reversing the max-min operation in (2.4). From this estimate of $\lambda(1-G(t))$, $t \geq 0$, one can obtain an estimate of $G(t)$, $t \geq 0$.

Simulated output times of $M/D/\infty$ and $M/M/\infty$ service systems have been used in calculation of the maximum likelihood estimators. Comparison of these estimates to the true service distribution and to another estimator is made in section 4.

3. PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATOR FOR $G(t)$

The maximum likelihood estimator of $\lambda \cdot G(t)$ is a step function with jumps at output times, $T_1 \leq T_2 \leq \ldots \leq T_n$. The first nonzero value or the estimate of $G(t)$ occurs at or after T_1, giving no information about $G(t)$ for $t < T_1$. This limitation may be removed by taking observations on N repeated runs of the service system starting empty. The ordered output times for all runs are used in calculating the estimator of $G(t)$. The maximum likelihood property of this estimator still exists and the estimator $\hat{G}_N(t)$ is given by,

$$\hat{G}_N(t) = \min[\hat{\lambda}(t)/N\lambda, 1], \tag{3.3}$$

where N is the number of runs.

A lower bound on the expected time of the first output from N runs is the expected first input time in N runs $1/\lambda N$. Extreme value theory suggests that asymptotically the time of the first observations on $G(t)$ will become smaller as the number of sums increases. Let D_{ij} be the departure time of the ith customer in the jth run where $i=1, 2, \ldots n$, $j=1, 2, \ldots N$. $D_{ij}=T_{ij}+S_{ij}$ where T_{ij} and S_{ij} are arrival and service, respectively. The first departure over N runs will take place at time given by $\min_{i\&j} \{T_{ij}+S_{ij}\}$. By extreme value theory, $\min_{i\&j} \{T_{ij}+S_{ij}\}$ will have Weibull distribution asymptotically, Gumbel [3], no matter what may be the distribution function for the random variable $(T_{ij}+S_{ij})$

provided $(T_{ij}+S_{ij})>0$. The expected value of a random variable x having Weibull distribution is given by

$$E(x)=\alpha\cdot\Gamma(1+1/\beta), \tag{3.4}$$

where α is the scale parameter and β is the shape parameter for the Weibull distribution. The scale parameter α can be estimated as the mth order statistic (m counted from bottom) for which

$$1-(m/N+1)=1/e=1/2.718\ \ldots\ . \tag{3.5}$$

As N increases m will increase which means the value of α, the scale parameter, will decrease. Thus the expected value given by Equation (3.4) will go to zero asymptotically for large values of N. In the context of the service system, this will mean that the expected time of first departure will asymptotically decrease to the lower support of the distribution as the number of runs increase. In other words the mean of the minimum order statistic of a random variable is of the order of the quantile for which the probability value is $\left(\frac{1}{N+1}\right)$ and thus will decrease to the smallest possible value of the random variable asymptotically with N.

A simple illustration can be given by considering service time to be a constant, t_0. The expected time of the first departure in N runs, each run with n observations is given by

$$\begin{aligned} E\min_{i\&j}\{T_{ij}+S_{ij}\} &= E\{\min(T_{ij})+t_0\} \\ &= E\{\min(T_{ij})\}+t_0 \\ &= (1/\lambda Nn)+t_0, \end{aligned} \tag{3.6}$$

where λ is the arrival rate of a Poisson arrival stream.

It can be seen from (3.6) that the expected time for the first departure in case of constant service time t_0 converges to the lower bound of the support of the service time distribution faster than $1/(N+1)$ as long as $n>1$.

The proof by Marshall and Proschan [4] of strong consistency of the maximum likelihood estimate of a distribution function under the assumption of increasing failure rate may be applied to show that the maximum likelihood estimator of an increasing distribution function $G(t)$ is strongly consistent at the points of continuity; i.e.,

$$\hat{G}_N(t)=G(t), \tag{3.1}$$

with probability 1, for a sufficiently large number of repeated observations N on the service system output starting empty. This may be done because the failure rate function $r(t)$ given as

$$r(t)=\frac{f(t)}{1-F(t)},$$

for an increasing failure rate distribution function $F(t)$ corresponds to a nondecreasing intensity function $\lambda(t)$ of a nonstationary Poisson process. In fact $\lambda(t)$ is the failure rate function of the distribution function of the event times conditional on previous event times. The maximum likelihood estimate of $\lambda(t)$ based on event times of a nonstationary, nondecreasing intensity Poisson process is the same as the maximum likelihood estimator of the failure rate function $r(t)$ from a nondecreasing failure rate distribution.

4. NUMERICAL RESULTS AND COMPARISON WITH BROWN'S [5] ESTIMATOR

4.1 Brown's Estimator

In Figure 1, the number of customers in the system $N(t)$ is plotted against time t. Let the origin on the time axis be shifted to the right such that it coincides with the first output after the old origin 0.

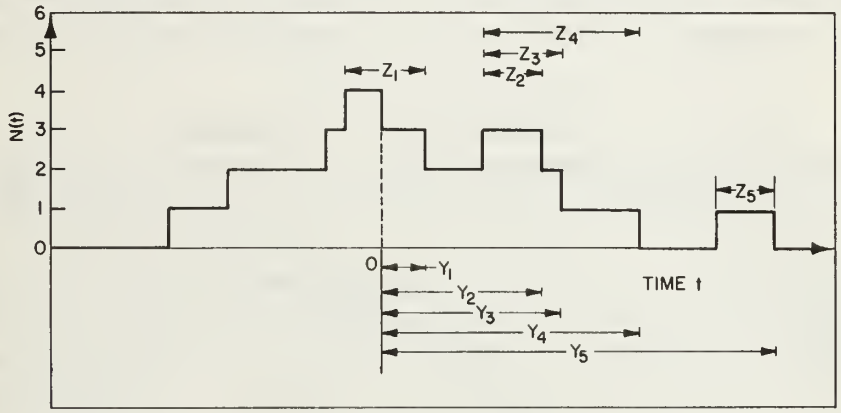

FIGURE 1. Number of units in the system $N(t)$ at time 't' vs. time t.

Y_i, $i=1, 2, \ldots, n$ is the time between the new origin $0'$ and the ith output point after the new origin. Z_i is the time from Y_i to the nearest input point prior to Y_i. For a stationary input process and independent identically distributed service times in steady state behavior, the Z_i are independent and identically distributed. Let $H(\cdot)$ be the distribution function of Z_i, $i=1, 2, \ldots, n$ and $\hat{H}_n(\cdot)$ the empirical distribution function based on observations $Z_1, Z_2, \ldots, Z_n$.

Then,

$$Pr[Z_i > x] = Pr[\text{time back from } i\text{th output to last previous input} > x]$$

$$= (1-H(x))$$

$$= Pr[\text{no input in the interval of length } x \cap \text{service takes longer than } x]$$

$$= e^{-\lambda x} \cdot (1-G(x))$$

or $1-H(x) = e^{-\lambda x}(1-G(x))$.

Thus, an estimate of $H(\cdot)$, given by $H_n(x)$, may be used for estimating the service distribution function $G(x)$:

$$\hat{G}(x) = 1 - e^{\lambda x}(1 - \hat{H}_n(x)). \tag{4.1}$$

This estimate may not be nondecreasing. A nondecreasing estimate of $G(x)$ is obtained by modifying Equation (4.1),

$$\hat{G}_n(x) = \max[0, \max_{i=1,2,\ldots n\hat{H}_n(x)} \{1 - e^{\lambda Z_i}(1 - \hat{H}_n(Z_i))\}]. \tag{4.2}$$

4.2 Numerical Results

The maximum likelihood estimate of the hidden service distribution function $G(x)$ was obtained for simulated operation of an $M/M/\infty$ system for various arrival rates λ and service rates μ. The results shown in Figure 2 correspond to an arrival rate of $\lambda = 1$ customer/min and exponential service at rate

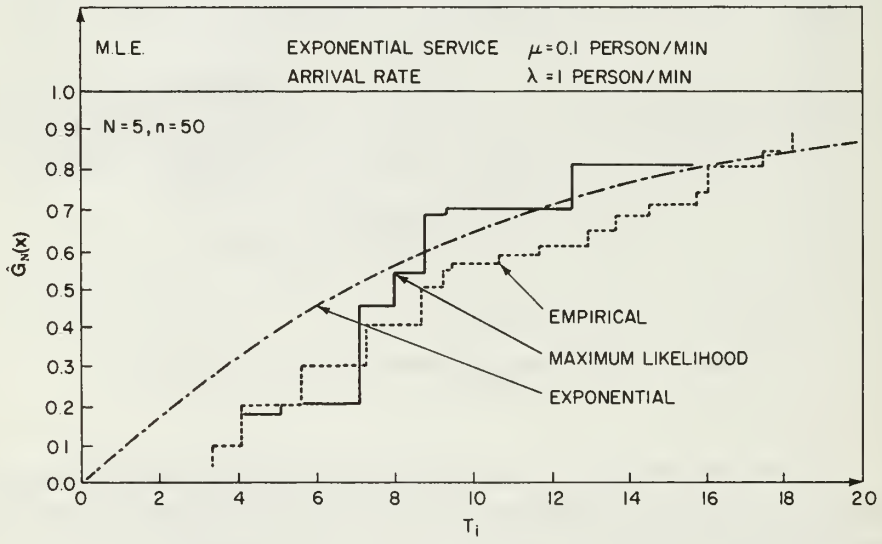

FIGURE 2. Maximum Likelihood Estimate (MLE).

$\mu = 0.1$ customer/min. Simulation was carried out for five runs, each consisting of 50 observations. The estimated values $\hat{G}_N(x)$ are plotted against output times T_i for $i = 1, 2, \ldots, n$. The empirical distribution function as well as an exponential distribution function for the service times are also plotted for the purpose of comparing the simulated results. It can be seen that the estimated distribution function is close to the empirical and the actual distribution function. Brown's estimator was simulated for an $M/M/\infty$ system in steady state with $\lambda = 1$ customer/min. and exponential service at rate $\mu = 0.5$ customer/min. Results are shown in Figure 3 and it is found that Brown's estimated distribution function is close the empirical as well as actual distribution function.

Further simulations with different exponential service rates has shown that while Brown's method gives reasonable results for a system having service rate close to or larger than arrival rate the maximum likelihood estimator is useful for slow service rate systems having large numbers of customers in the

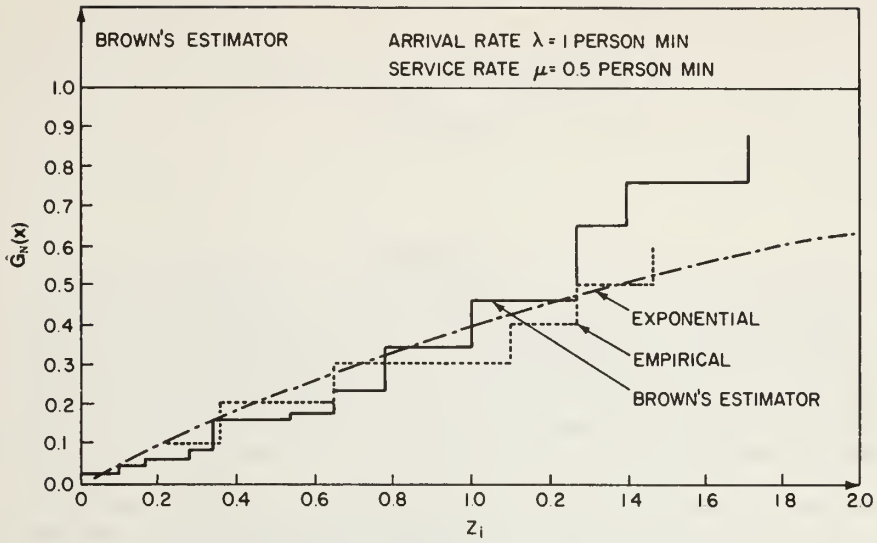

FIGURE 3. Brown's estimator.

system. This contrasting behaviour of the two estimators may be used in order to obtain better results by using Brown's estimator in case of fast service and maximum likelihood estimator in case of slow service.

Simulation of estimators was also performed for constant service times. The same remarks as above apply to the usefulness of the two estimators relative to service rate. It was also noted that the maximum likelihood estimator converged to the true unknown service time from above while Brown's estimator was less biased.

REFERENCES

[1] Boswell, M. T., "Estimation and Testing Trend in a Stochastic Process of Poisson Type," The Annals of Mathematical Statistics, *37*, 1564–1573 (1966).

[2] Brown, M., "An Estimation Problem in $M/G/\infty$ Queues with Applications to Traffic," Technical Rept. No. 59, Department of Operations Research (Cornell University, Ithaca, New York, 1968).

[3] Gumbel, E. J., *Statistics of Extremes* (Columbia University Press, New York, 1968).

[4] Marshall, A. W. and F. Proschan, "Maximum Likelihood Estimation for Distributions with Monotone Failure Rate," The Annals of Mathematical Statistics *36*, 69–77 (1965).

[5] Mirasol, N. M., "The Output of an $M/G/\infty$ Queuing System is Poisson," Operations Research *11*, 282–284 (1963).

THE SINGLE SERVER QUEUE IN DISCRETE TIME-NUMERICAL ANALYSIS III

Marcel F. Neuts* and Eugene Klimko

Purdue University

ABSTRACT

This paper deals with the *stationary analysis* of the finite, single server queue in discrete time. The following *stationary* distributions and other quantities of practical interest are investigated: (1) the joint density of the queue length and the residual service time, (2) the queue length distribution and its mean, (3) the distribution of the residual service time and its mean, (4) the distribution and the expected value of the number of customers lost per unit of time due to saturation of the waiting capacity, (5) the distribution and the mean of the waiting time, (6) the asymptotic distribution of the queue length following departures.

The latter distribution is particularly noteworthy, in view of the substantial difference which exists, *in general*, between the distributions of the queue lengths at arbitrary points of time and those immediately following departures.

1. INTRODUCTION

This paper is a direct sequel to [2], to which we refer for a detailed definition and for the assumptions of the finite, discrete time queue. For easy reference, we only give a summary of the notation here.

NOTATION

L_1 Maximum number of customers allowed in the system at any time. All excess customers are lost and do not return.

L_2 Maximum duration of the service time of a single customer.

r_j Probability that a service lasts for j units of time, $j=1, \ldots, L_2$. We assume without loss of generality that $r_{L_2}>0$. Also $r_1+\ldots+r_{L_2}=1$.

K Maximum number of arrivals during a unit of time. It is assumed that $K<L_1$.

p_j Probability that j customers arrive during a unit of time, $j=0, 1, \ldots, K$. We assume without loss of generality that $p_0>0$. and $p_K>0$. Also $p_0+\ldots+p_K=1$.

X_n The number of customers in the system at time $n+$.

Y_n The number of time units until the customer in service at time $n+$ completes service. We note that $0\leqslant Y_n\leqslant L_2$ and that $Y_n=0$ if and only if $X_n=0$.

In [2], it was shown that the bivariate sequence $\{(X_n, Y_n), n\geqslant 0\}$ is an irreducible, aperiodic Markov chain with state space $\{(0, 0)\}\cup\{(1, 2, \ldots, L_1)\times(1, \ldots, L_2)\}$. Its transient behavior was discussed and investigated numerically in [2]. In this paper we first discuss the *stationary* joint distribution of the queue length X_n and the residual service time Y_n.

*The research of this author was supported by the National Science Foundation, Contract No. GP 28650

2. THE EQUATIONS FOR THE STATIONARY JOINT PROBABILITIES OF X_n AND Y_n.

We denote the stationary probabilities by $P(i, j)$ for $i = 1, \ldots, L_1$ and $j = 1, \ldots, L_2$; $P(0, 0)$ is the stationary probability that the queue is empty. The stationary joint density of X_n and Y_n is the unique solution to the following system of linear equations.

(1) a. $$P(0, 0) = p_0[P(1, 1) + P(0, 0)].$$

b. $$P(i, j) = \sum_{\nu=0}^{i} p_{i-\nu} P(\nu, j+1) + r_j r_{L_2}^{-1} P(i, L_2),$$

for $1 \leq i \leq K$, $1 \leq j \leq L_2 - 1$.

c. $$P(i, j) = \sum_{\nu=i-K}^{i} p_{i-\nu} P(\nu, j+1) + r_j r_{L_2}^{-1} P(i, L_2),$$

for $K+1 \leq i \leq L_1 - 1$, $1 \leq j \leq L_2 - 1$.

d. $$P(L_1, j) = P(L_1, j+1) + \sum_{\nu=1}^{K} \left(1 - \sum_{k=0}^{\nu-1} p_k\right) P(L_1 - \nu, j+1) + r_j r_{L_2}^{-1} P(L_1, L_2),$$

for $1 \leq j \leq L_2 - 1$.

e. $$P(i, L_2) = r_{L_2} \left\{ \frac{p_0 p_i}{1 - p_0} P(1, 1) + \sum_{\nu=0}^{i} p_{i-\nu+1} P(\nu, 1) \right\},$$

for $1 \leq i \leq K$.

f. $$P(i, L_2) = r_{L_2} \sum_{\nu=i-K+1}^{i+1} p_{i-\nu+1} P(\nu, 1),$$

for $K+1 \leq i \leq L_1 - 1$.

g. $$P(L_1, L_2) = r_{L_2} \sum_{\nu=1}^{K} \left(1 - \sum_{k=0}^{\nu-1} p_k\right) P(L_1 - \nu + 1, 1),$$

h. $$P(0, 0) + \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} P(i, j) = 1.$$

The system (1) contains $L_1 L_2 + 1$ *independent* linear equations in $L_1 L_2 + 1$ unknowns. We shall show that its solution may be conveniently expressed in terms of the solution of a homogeneous system of

L_1 equations in L_1 unknowns. Moreover, the latter system has a particular structure which greatly simplifies its numerical solution.

We denote by $\underline{P}_j$ the L_1-tuple $[P(1, j), \ldots, P(L_1, j)]$ for $j=1, \ldots, L_2$. We also introduce the $L_1 \times L_1$ matrices A and B defined as follows:

$$A = \begin{pmatrix}
p_0 & p_1 & p_2 & \ldots & p_K & 0 & \ldots & 0 & 0 & 0 \\
0 & p_0 & p_1 & \ldots & p_{K-1} & p_K & \ldots & 0 & 0 & 0 \\
0 & 0 & p_0 & \ldots & p_{K-2} & p_{K-1} & \ldots & 0 & 0 & 0 \\
0 & 0 & 0 & \ldots & p_{K-3} & p_{K-2} & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
 & & & & & & & p_{K-2} & p_{K-1} & p_K \\
 & & & & & & & p_{K-3} & p_{K-2} & p_{K-1}+p_K \\
 & & & & & & & \vdots & \vdots & \vdots \\
 & & & & & & & \vdots & \vdots & \vdots \\
 & & & & & & & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & 0 & \ldots & p_1 & p_2 & 1-p_0-p_1-p_2 \\
0 & 0 & 0 & \ldots & 0 & 0 & \ldots & p_0 & p_1 & 1-p_0-p_1 \\
0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & p_0 & 1-p_0 \\
0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & 1
\end{pmatrix}$$

$$B = \begin{pmatrix}
\frac{p_1}{1-p_0} & \frac{p_2}{1-p_0} & \frac{p_3}{1-p_0} & \ldots & \frac{p_k}{1-p_0} & 0 & \ldots & 0 & 0 & 0 \\
p_0 & p_1 & p_2 & \ldots & p_{K-1} & p_K & \ldots & 0 & 0 & 0 \\
0 & p_0 & p_1 & \ldots & p_{K-2} & p_{K-1} & \ldots & 0 & 0 & 0 \\
0 & 0 & p_0 & \ldots & p_{K-3} & p_{K-2} & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
 & & & & & & & p_{K-2} & p_{K-1} & p_K \\
 & & & & & & & p_{K-3} & p_{K-2} & p_{K-1}+p_K \\
 & & & & & & & \vdots & \vdots & \vdots \\
 & & & & & & & \vdots & \vdots & \vdots \\
 & & & & & & & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & 0 & \ldots & p_2 & p_3 & 1-p_0-\ldots-p_3 \\
0 & 0 & 0 & \ldots & 0 & 0 & \ldots & p_1 & p_2 & 1-p_0-p_1-p_2 \\
0 & 0 & 0 & \ldots & 0 & 0 & \ldots & p_0 & p_1 & 1-p_0-p_1 \\
0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & p_0 & 1-p_0
\end{pmatrix}$$

In terms of A and B, the equations $(1b-g)$ may be written as

(2)
$$\underline{P}_j = \underline{P}_{j+1}A + r_j r_{L_2}^{-1}\underline{P}_{L_2}, \qquad 1 \leq j \leq L_2 - 1,$$

$$\underline{P}_{L_2} = r_{L_2}\underline{P}_1 B.$$

The latter system is equivalent to the equations

(3)
$$\underline{P}_j = r_{L_2}^{-1}\underline{P}_{L_2}\sum_{\nu=j}^{L_2} r_\nu A^{\nu-j}, \qquad 1 \leq j \leq L_2 - 1,$$

$$\underline{P}_{L_2} = \underline{P}_{L_2}\left(\sum_{\nu=1}^{L_2} r_\nu A^{\nu-1}\right)B.$$

We now observe that both A and B are stochastic matrices, that A is upper triangular and that the matrix B has only one subdiagonal. We shall say, for brevity, that B is *nearly upper triangular*. Since $r_1 + \ldots + r_{L_2} = 1$, and A is an upper triangular stochastic matrix, the matrix $\sum_{\nu=1}^{L_2} r_\nu A^{\nu-1}$ is stochastic and upper triangular. The stochastic matrix B is irreducible, so that the matrix

(4)
$$Q = \sum_{\nu=1}^{L_2} r_\nu A^{\nu-1}B,$$

is irreducible and stochastic. Finally it is easy to verify that Q is nearly upper triangular.

The vector $\underline{P}_{L_2}$ is therefore proportional to the vector of the stationary probabilities of the matrix Q. The nearly upper triangular form of the matrix Q makes the numerical computation of the vector $\underline{P}_{L_2}$—up to a positive multiplicative constant—particularly simple. The vector $\underline{P}_{L_2}$ is proportional to the vector $(t_1, t_2, \ldots, t_{L_1})$, whose components may be computed recursively as follows

(5)
$$t_1 = 1,$$

$$t_2 = (1 - q_{11})q_{21}^{-1},$$

$$t_k = q_{k,k-1}^{-1}\left[t_{k-1}(1 - q_{k-1,k-1}) - \sum_{\nu=1}^{k-2} t_\nu q_{\nu,k-1}\right], \qquad 3 \leq k \leq L_1.$$

It is easy to verify that none of the entries $q_{k,k-1}$, $2 \leq k \leq L_1$ vanish, so that by using the first equation in (2), the vectors $\underline{P}_j$, $j = 1, \ldots, L_2 - 1$, may be computed up to a common, positive multiplicative constant. Equation (1a) is then used to determine $P(0, 0)$ up to the same multiplicative constant. This constant may finally be computed using Equation (1h). The stationary joint density of the queue length and the residual service time is therefore determined.

3. THE STATIONARY DENSITY OF THE WAITING TIME

The support of the stationary density $\{w_j\}$ of the waiting time consists of the integers $0, 1, \ldots, L_1L_2$. Clearly $w_0 = P(0, 0)$ and for $1 \leq j \leq L_1L_2$, the density may be written symbolically as the convolution polynomial

$$\{w_j\} = P(1, \cdot) + P(2, \cdot) * \{r_\nu\} + P(3, \cdot) * \{r_\nu\}^{(2)} + \ldots + P(L_1, \cdot) * \{r_\nu\}^{(L_1 - 1)}, \tag{6}$$

where $\{r_\nu\}$ is the density of the service time.

The numerical computation of the w_j, $1 \leq j \leq L_1L_2$, by using a convolution analogue of *Horner's algorithm* for polynomials was discussed in [2].

4. THE STATIONARY DENSITY OF THE NUMBER OF LOST CUSTOMERS PER UNIT OF TIME

Since the waiting room is finite, it is possible that customers will be lost due to the waiting room being full at their arrival time. It is therefore of interest to know the stationary density $\{\varphi_j\}$ of the number of lost customers per unit of time. It has its support on the integers $0, 1, \ldots K$ and may be determined by the explicit expressions

$$\varphi_j = \sum_{k=j}^{K} p_k \sum_{\nu=1}^{L_2} P(L_1 - k + j, \nu), \qquad 1 \leq j \leq K, \tag{7}$$

$$\varphi_0 = 1 - \sum_{j=1}^{K} \varphi_j.$$

Knowing the joint density discussed in section 2, the probabilities $\{\varphi_j\}$ are readily computed.

5. THE STATIONARY DENSITY OF THE QUEUE LENGTH AT DEPARTURES

The probabilities associated with the queue length at departure times, are primarily of interest in the analytic treatment of queues of $M|G|1$ type. Although they are frequently examined, their inherent applied interest is limited.

As we shall indicate below, the density of the queue length following departures may easily be obtained from auxiliary quantities which are computed in the process of evaluating the joint stationary density, discussed in section 2. In view of the importance ascribed to this density in the applied queueing literature, we decided to investigate its computational aspects. Note the very substantial difference which may exist between it and the stationary density of X_n.

The queue lengths following departures form an irreducible, aperiodic Markov chain with state space $\{0, 1, \ldots, L_1 - 1\}$. Let us denote its transition probability matrix by T. Furthermore, let $\theta_k(i, \nu)$ be the probability that in k consecutive units of time during which no departures occur, ν customers join the queue, given that the queue length at the beginning of the first unit of time was i.

The entries of T are then given by

$$T_{0j} = \sum_{k=1}^{L_2} r_k \sum_{h=1}^{K} p_h (1 - p_0)^{-1} \theta_k(h, j - h + 1), \qquad \text{for } 0 \leq j \leq L_1 - 1, \tag{8}$$

$$T_{ij}=\sum_{k=1}^{L_2} r_k\theta_k(i,j-i+1), \qquad \text{for } 1\leq i\leq j+1,$$

$$T_{ij}=0, \qquad \text{for } i>j+1.$$

We note that the transition probability matrix T is *nearly upper triangular*. The stationary probabilities corresponding to T may be calculated by a simple recursion such as in Formula (5). In order to evaluate the entries of the matrix T, we first show that

$$\theta_k(i,j-i+1)=(A^k)_{i,j+1}, \qquad \text{for } 1\leq i\leq L_1,\ 0\leq j\leq L_1-1, \tag{9}$$

where A is the upper triangular matrix defined in section 2.

For $k=1$, we find that

$$\theta_1(i,j-i+1)=p_{j-i+1}, \qquad \text{for } 0\leq j-i+1\leq K,\ j\leq L_1-2, \tag{10}$$

$$=\sum_{\nu=L_1-i}^{K} p_\nu, \qquad \text{for } L_1-K\leq i\leq L_1,\ j=L_1-1,$$

$$=0, \qquad \text{for all other pairs } (i,j),$$

so that Equation (8) holds for $k=1$. Furthermore

$$\theta_{k+1}(i,j-i+1)=\sum_{\nu=\max(0,j-K)}^{j}\theta_k(i,\nu-i+1)p_{j-\nu}, \tag{11}$$

for $0\leq j\leq L_1-2$, and

$$\theta_{k+1}(i,L_1-i)=\sum_{\nu=L_1-K}^{L_1}\theta_k(i,\nu-i)\sum_{h=L_1-\nu}^{K}p_h,$$

for $1\leq i\leq L_1$. When expressed in terms of the matrix A, Formula (11) proves (9) inductively.

The matrix T can be compactly written as

$$T=C\sum_{k=1}^{L_2} r_kA^k, \tag{12}$$

where $C_{ij}=p_j(1-p_0)^{-1}$ for $1\leq j\leq K$; $C_{i-1,i}=1$, for $2\leq i\leq L_1$, and $C_{ij}=0$, for all other pairs (i,j).

The relation between the limiting distribution of the queue length following the nth departure and the stationary queue length distribution is noteworthy. A well-known theorem, from Reference [3], states that in a stable $M|G|1$ queue with single arrivals, the queue length at time t and the queue length following the nth departure have *the same limiting distribution* as t and n respectively tend to infinity.

An analogous result holds for the [illegible] discussed by Dafermos and Neuts [1], provided that the probability of two or more [illegible] unit of time is zero. This result is proved by an exact analogue of the argument [illegible], so we shall omit the proof.

In the case of group arrivals ($K > 2$) [illegible] between those two limiting distributions. Theorems which relate those [illegible] using the theory of Markov renewal processes, but the resulting formulas are [illegible]. We shall not pursue this topic here, but we offer as an illustration some numerical [illegible] which has rare arrivals of large groups of customers.

We considered a queue with $L = 100$ [illegible], and with $p_0 = 0.975$, $p_{20} = 0.025$, $r_1 = r_2 = 0.5$. Although the traffic intensity ρ for the [illegible] 0.975, examination of the transient behavior shows that this queue converges *very* [illegible] to its stationary phase.

The limiting distribution of the queue [illegible] *following a departure* has a mean equal to 32.2864. In contrast, the limiting distribution of the [illegible] *at time n* has a mean equal to 24.1752. In addition, we list a summary of the numerical [illegible] stationary distributions. π_k is the stationary probability of at most k customers at [illegible] stationary probability of at most k customers, following a departure from the system (see [illegible]).

[illegible]

k	π_k	p_k^*
0	[illegible]	0.0249
20	[illegible]	0.1230
40	[illegible]	0.4799
60	[illegible]	0.8459
80	[illegible]	[illegible]
90	[illegible]	[illegible]
100	[illegible]	[illegible]
110	[illegible]	0.9998
120	[illegible]	0.9999
130	[illegible]	[illegible]
140	[illegible]	[illegible]

The greater limiting probability [illegible] following departures may appear to be paradoxical at a casual reading. A moment's reflection shows however that, on the contrary, this is to be anticipated in stable queues with group [illegible]. In our example, the queue length will typically be zero for long intervals of time because [illegible] value of p_0. The averaging procedure involved in the stationary distribution of the queue length [illegible] n heavily favors the lower values of k. The limiting distribution of the queue length following the nth departure effectively ignores the long idle periods and results primarily from the behavior of the queue during the service of the large groups of customers. The high probabilities of the [illegible] values of k in this distribution are therefore not surprising.

This example strikingly shows that the [illegible] distribution of the queue features may be of limited practical value, even in very stable [illegible]. Most realizations of the queue length process in our example will exhibit very substantial [illegible], which are not reflected in the asymptotic distri-

butions. The practical questions related to queues of this type can only be answered after analyzing their *transient* behavior. The exclusive concern with asymptotic results in "practical" discussions of queueing theory is therefore regrettable.

6. COMPUTATIONAL ORGANIZATION

In order to minimize both the computation time and the required memory storage, we took advantage of the highly structured form of the matrices Q and T in Equations (4) and (11), respectively.

The basic matrix is the upper triangular polynomial matrix $Q^*=\{q^*_{ij}\}$

$$Q^*=\sum_{\nu=1}^{L_2} r_\nu A^{\nu-1}. \tag{13}$$

The rows of this matrix are similar in the sense that

$$q^*_{i,\,i+\nu}=q^*_{1,\,\nu+1} \tag{14}$$

$$\text{for } \nu=0,1,2,\ \ldots\ L_1-i-1;\ i=2,3,\ \ldots,L_1-1.$$

Furthermore the matrix Q^* is stochastic, so that

$$q^*_{i,\,L_1}=1-\sum_{j=i}^{L_1-1} q^*_{ij}. \tag{15}$$

Therefore, the first row determines the entire matrix. This permits the storage of Q^* using only L_1 memory spaces, rather than the $(L_1^2+L_1)/2$ spaces required for an arbitrary upper triangular matrix. The resulting saving in memory space is substantial for large queues and in fact makes the analysis of queue lengths up to 800 feasible. Computation of the matrix Q^* is performed by using Horner's method for the formation of polynomials, i.e., by recursive computation as follows

$$Q_1^*=(r_{L_2}A+r_{L_2-1}I) \tag{16}$$

$$Q_n^*=Q_{n-1}^*A+r_{L_2-n}I \qquad n=2,\ \ldots,\ L_2-1.$$

Each of the successive matrices Q_n^* is completely determined by its top row. The right-most elements are not needed and therefore are not computed. The top row entries of Q_n^* are rapidly calculated by means of the formulas

$$q_{1j}^{*(n+1)}=\sum_{i=0}^{\min(K,\,j)} p_i q_{j-i}^{*(n)}+r_{L_2-n}p_j, \quad \text{for } j\leq k \tag{17}$$

$$q_{1j}^{*(n+1)}=\sum_{i=0}^{\min(K,\,j)} p_i q_{j-i}^{*(n)}, \qquad \text{for } K<j\leq L_1.$$

The matrix Q has the form

$$\begin{pmatrix}
q_{11} & q_{12} & q_{13} & q_{14} & \cdots & q_{1,L_1-4} & q_{1,L_1-3} & q_{1,L_1-1} & q_{1L_1} \\
q_{21} & q_{22} & q_{23} & q_{24} & \cdots & q_{2,L_1-3} & q_{2,L_1-2} & q_{2,L_1-1} & q_{2L_1} \\
0 & q_{21} & q_{22} & q_{23} & \cdots & q_{2,L_1-4} & q_{2,L_1-3} & q_{3,L_1-1} & q_{3,L_1} \\
0 & 0 & q_{21} & q_{22} & \cdots & q_{2,L_1-5} & q_{2,L_1-4} & q_{4,L_1-1} & q_{4,L_1} \\
\cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & \cdot & \cdot & \cdots & q_{21} & q_{22} & q_{L_1-2,L_1-1} & q_{L_1-2,L_1} \\
0 & 0 & \cdot & \cdot & \cdots & 0 & q_{21} & q_{L_1-1,L_1-1} & q_{L_1-1,L_1} \\
0 & 0 & \cdot & \cdot & \cdots & 0 & 0 & q_{L_1,L_1-1} & q_{L_1,L_1}
\end{pmatrix}$$

where the third through the last rows, except for the last two columns, are essentially repetitions of the second row. The last column is determined by the condition that the rows sum to one. We therefore need to compute and store only the first and second rows and the (L_1-1)-st column. This requires $3L_1-4$ memory cells for the storage of the Q matrix rather than the $L_1+(L_1+2)(L_1-1)/2$ required for an arbitrary nearly upper triangular matrix. The top row elements of Q are given by

$$q_{1j}=\frac{p_j}{1-p_0}q^*_{11}+\sum_{i=0}^{\min(K,j-1)}p_i q^*_{1,j-i+1}, \qquad \text{for } j\leq K \tag{18}$$

$$=\sum_{i=0}^{K}p_i q^*_{1,j-i+1}, \qquad \text{for } K<j<L_1-1$$

$$q_{1,L_1-1}=p_0\left(1-\sum_{j=1}^{L_1-1}q^*_{1j}\right)+\sum_{i=1}^{K}p_i q^*_{1,L_1-i};$$

the second row elements of the Q matrix are

$$q_{2j}=\sum_{i=0}^{\min(K,j-1)}p_i q^*_{1,j-1}, \qquad \text{for } 1\leq j\leq L_1-2; \tag{19}$$

and the (L_1-1)-st column elements are calculated by using

$$q_{i,L_1-1}=p_0\left(1-\sum_{j=1}^{L_1-i}q^*_{1j}\right)+\sum_{k=1}^{\min(K,L_1-i)}p_k q^*_{1,K_1-k}, \tag{20}$$

for $2\leq i\leq L_1$.

The stationary probabilities of the Q matrix were determined using Formula (5) and its compact representation by Formulas (18)–(20). For this purpose, a subroutine called STAPROB was written. The resulting stationary probability vector was identified temporarily with the vector $\underline{P}_{L_2}$. The vectors

$\underline{P}_{L_2}, \ldots, \underline{P}_1$ were successively obtained by (21). To execute this computation essentially only the top row of the matrix A is needed. The multiplications involved in

(21) $$P(i, j) = \sum_{k=1}^{\min(K, L_1+L_2-j)} a_k P(i+k-1, j+1) + a_0 P(i, L_2),$$

for $j = L_2 - 1, \ldots, 1$.

Finally $P(0, 0)$ is computed and $\underline{P}_0$ is normalized to satisfy the normalization condition $(1h)$.

The waiting-time distribution was calculated by a subroutine called WAIT. This subroutine was adapted from the program discussed in [2]. In cases where L_1L_2 is large, one may wish to print only the percentage points of the waiting-time distribution. A routine to do this was also written.

The computational procedure for the queue length following a departure is similar to that for the stationary queue length distribution. The polynomial

(22) $$\sum_{j=0}^{K} \ldots Q^j$$

is first computed and then the matrix T is determined. It is represented in a manner similar to that of the matrix Q. Only a modicum of additional computation is involved. The stationary distribution is then calculated by the subroutine STAPROB.

Testing

In addition to testing the program for its correctness, we compared the stationary probabilities with the transient probabilities after 60 time units. The latter were obtained by the methods developed in [2].

Computational Experience

Practical limits on the problem size were determined by the memory requirements. The available memory space of 150K octal required that $L_1 \leq 800$ approximately. This permits, for instance, queue lengths of size 800 with service times possibly up to 25 points. For problems of this magnitude the computation time was a limiting factor only in the determination of the waitingtime distribution. We ran examples, both with and without the determination of the waitingtime. The central processing times on the CDC 6500 at Purdue University for these examples are shown in Table 2. T_1 and T_2 are the actual program running times in seconds (without compilation and loading times), respectively, with and without the computation of the waitingtime distribution. For the example with $L_1 = 800$, $L_2 = 25$, $K = 4$, the time T_1 was in excess of 3,000 seconds and the computations were not completed even then. In all the examples, we used the same arrival distribution $p_0 = 0.8$, $p_1 = p_2 = p_3 = p_4 = 0.05$. The service time distribution for the first three examples was $r_1 = 0.075$, $r_2 = r_3 = r_4 = 0.05$, and $r_5 = 0.175$. In the last example, the service time distribution was a truncated geometric with $p = 0.5$ and the residual probability was added to r_{25}.

TABLE 2

L_1	L_2	K	T_1	T_2
100	5	4	5.751	0.945
200	5	4	22.539	2.221
400	5	4	69.774	6.612
800	25	4	> 3,071.032	26.290

7. CONCLUSIONS

Large discrete, single server queues in the stationary phase may be analyzed numerically. As we have shown, most queue features of interest, with the possible exception of the stationary waiting-time distribution, can be computed without the use of excessive processing times. This should be contrasted with simulation methods which are inherently ill-suited for the study of the stationary phase.

The prohibitive processing times required for the waiting-time distribution in large queues, raise the interesting question of how to evolve efficient numerical procedures for the evaluation of expressions of the general type

$$\sum_{i=1}^{n} A_i(\cdot) * F^{(i)}(\cdot)$$

which appear frequently in stochastic models of varied applied interest.

Finally, the example discussed in section 5, shows that in queues exhibiting large fluctuations, it may be hazardous to base conclusions on a single stationary distribution. In such cases one should study the transient behavior, whenever possible.

For further information on the algorithms discussed in this paper, one may contact either of the authors at the Department of Statistics, Purdue University, West Lafayette, Ind. 47907.

REFERENCES

[1] Dafermos, S. and M. F. Neuts, "A Single Server Queue in Discrete Time," Cahiers du Centre de Recherche Opérationnelle *13*, 23–40 (1971).

[2] Neuts, M. F., "The Single Server Queue in Discrete Time—Numerical Analysis I," Nav. Res. Log. Quarterly, *20*, 297–304 (1973).

[3] Takàcs, L., *Introduction to the Theory of Queues* (Oxford University Press, New York, 1962).

SOME EXPERIMENTS IN GLOBAL OPTIMIZATION

James K. Hartman

Naval Postgraduate School
Monterey, California

ABSTRACT

When applied to a problem which has more than one local optimal solution, most nonlinear programming algorithms will terminate with the first local solution found. Several methods have been suggested for extending the search to find the global optimum of such a nonlinear program. In this report we present the results of some numerical experiments designed to compare the performance of various strategies for finding the global solution.

I. INTRODUCTION

It is frequently the case in applied optimization studies that an algorithm which is known to converge to a global optimal solution under certain conditions (such as convexity) will be applied to a problem which does not satisfy these conditions. In particular, optimization problems which are suspected of having several local optima in addition to the global optimum are often solved using algorithms which will stop and indicate a solution whenever any local optimum is reached. In such cases a useful strategy is to repeat the solution process several times starting from different initial points to avoid accepting a solution which is only a local optimum. This is probably the most frequently suggested strategy for avoiding local solutions.

There are also other strategies for avoiding the local solutions in favor of the global optimum. This paper describes some numerical experiments which were done to compare the performance of several strategies for organizing such a global optimization.

II. THE PROBLEM

In order to develop and test strategies for avoiding local solutions it is necessary to specify a class of optimization problems to be considered. This paper will concentrate on the "essentially unconstrained" nonlinear programming problem

$$\text{(1)} \qquad \text{minimize } f(x)$$

$$\text{subject to} \qquad x \epsilon S \subseteq E^n,$$

where the local and global optimal solutions to (1) are known to occur in the interior of the set S. In such a problem the feasible region S determines a domain to be searched for solutions, but the boundaries of S do not determine the solutions. In this sense problem (1) can be considered "essentially unconstrained." The simplest way to specify the set S is to place upper and lower bounds on each variable. Since each of the strategies to be considered will involve random selections of x, it is necessary to confine the search to a bounded region S. In addition, search strategies S5 and S6 will partition S into smaller regions; these two strategies can only be conveniently described for S determined by upper and lower bounds on the variables.

"Essentially unconstrained" problems arise frequently as the "unconstrained" subproblems in interior penalty function algorithms such as the Sequential Unconstrained Minimization Technique of Fiacco and McCormick [3]. In the SUMT method, if the original nonlinear program is not a convex program, then the subproblem (1) may have local solutions which are distinct from the global solution.

For problems like (1) a local optimal solution can be obtained by applying any of the efficient unconstrained descent algorithms (such as the Davidon-Fletcher-Powell method) to minimize the function $f(x)$ while being careful not to penetrate the boundary of S. We shall now consider several strategies which try to ensure that the local solution we finally accept is, in fact, a global minimum.

III. STRATEGIES FOR AVOIDING LOCAL SOLUTIONS

Six different strategies for organizing a global optimization are compared in this paper. These are briefly described below with references to more complete descriptions when they exist.

STRATEGY S1 (From the folklore):

a. Set $k=1$.

b. Let x^k be a vector chosen at random in the search region S. Starting at x^k perform an unconstrained minimization search on the function $f(x)$ terminating at the local minimum x^{k*}.

c. Replace k with $k+1$ and go to step b. At each stage retain the best local solution obtained to date.

S1 is the strategy suggested in section I. Intuitively the problem with this strategy is that it may repeatedly search to the same local minimum if the starting points x^k happen to be chosen within the "range of attraction" of that local minimum. The next three strategies attempt to solve this problem.

STRATEGY S2:

a. Set $k=1$. Let f^* be the objective function value for the best local solution so far obtained. Initially $f^*=+\infty$.

b. Randomly select points $x \epsilon S$ until one is found with $f(x)<f^*$. Call this point x^k.

c. Starting at x^k perform an unconstrained minimization search terminating at a new local minimum x^{k*}.

d. Set $f^*=f(x^{k*})$, replace k with $k+1$, and go to step b.

In S2 a minimization (step c) is initiated at x^k only if $f(x^k)$ is smaller than the best solution f^* found to date. Hence, each successive minimization gives a new local minimum which is better than any found so far. The same local minimum cannot be located twice. It is, however, much more difficult to determine the starting points x^k for strategy S2 than for S1.

STRATEGY S3 (Bocharov [1]):

a. Choose x^1 randomly in S. Set $k=1$.

b. Starting from x^k perform an unconstrained minimization terminating at the local minimum x^{k*}.

c. Choose a direction vector $d^k \epsilon E^n$ at random and consider $f(x^{k*}+\alpha d^k)$ as the positive scalar α increases. Moving away from x^{k*} in direction d^k, the funtion f must initially increase (since x^{k*} is a local minimum). Continue to increase α until f begins to decrease when $\alpha=\alpha^k$.

d. Let $x^{k+1}=x^{k*}+\alpha^k d^k$, replace k with $k+1$, and go to step b.

STRATEGY S4 (Bocharov[1]):

S4 is the same as S3 except that in step c, instead of choosing the direction at random, d^k is

chosen to be the direction of overall progress from the most recent minimization

$$d^k = x^{k*} - x^k. \tag{2}$$

Both S3 and S4 attempt to prevent repeated minimization to the same local optimum by moving out of the region of attraction of the most recent local solution before starting the next minimization. By continuing in the direction (2), Strategy S4 hopes to also avoid local minima detected before the most recent minimum.

Strategies S5 and S6 are considerably different from the first four methods. While S1–S4 attempt to choose good starting points for repeated local minimizations, S5 and S6 attempt to gain information about the entire search region S, gradually concentrating their attention on portions of S which are, in some sense, "likely" to contain the global minimum. S5 and S6 are most easily described for problems where S is determined by lower and upper bounds on each variable:

$$S = \{x \epsilon E^n \mid l_i \leq x_i \leq L_i, \qquad i = 1, \ldots, n\}.$$

For ease of presentation we will restrict our attention to such problems.

STRATEGY S5 (Piecewise Coordinate Projection—Zakharov [5]):

a. Set up an initially empty list of points, and let $\hat{S} = \{x \epsilon E^n \mid \hat{l}_i \leq x_i \leq \hat{L}_i, i = 1, \ldots, n\}$ be the "remaining feasible region." Let $\hat{S} = S$ initially.

b. Randomly choose N points $x^k \epsilon \hat{S}$, compute $f(x^k)$ for each, and adjoin them to the list.

c. For each coordinate x_i of $x (i = 1, \ldots, n)$ separate the remaining feasible interval $[\hat{l}_i, \hat{L}_i]$ into m equal subintervals. Let $X_{ij} = \{x^k$ in the list whose ith component is in the jth subinterval of $[\hat{l}_i, \hat{L}_i]\} = \{x^k \mid (j-1)(\hat{L}_i - \hat{l}_i)/m \leq x_i^k - \hat{l}_i < j(\hat{L}_i - \hat{l}_i)/m\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Then $X_{i1}, X_{i2}, \ldots, X_{im}$ describe the projection of the list of points x^k into the m subintervals of the ith coordinate axis.

d. By considering $\{f(x^k) \mid x^k \epsilon X_{ij}\} (i = 1, \ldots, n;\ j = 1, \ldots, m)$ select the subinterval set X_{st} which is considered most likely to contain the global minimum. Briefly, this is done by selecting the subinterval set for which the average functional value is smallest, being careful to avoid choices based on insufficient information (for more details see Zakhorov [5]).

e. By redefining $\hat{l}_s$ and $\hat{L}_s$ delete the subinterval sets X_{sj} $(J = 1, \ldots, m;\ j \neq t)$ from the remaining feasible region. Delete each point x^k in the list whose sth coordinate is in a deleted subinterval X_{sj}. Go to step b.

As the remaining feasible region $\hat{S}$ gradually shrinks, the global minimum will be more and more closely bracketed. The problem with this method is that the most promising subinterval must be determined on the basis of the sample of points x^k chosen so far. There is always a chance that a subinterval chosen for deletion will, in fact, contain the global minimum solution, and once it is deleted it can never be recovered.

Strategy S6 attempts to solve this problem by retaining the entire region S throughout and using a probabilistic allocation device to concentrate attention on areas in S which are most promising. This algorithm is new and is still under development. Initial results show some promise, but considerable improvement is still necessary.

STRATEGY S6 (Coordinatewise Allocation):

a. Define a marginal probability distribution function Φ_i on the feasible interval $[l_i, L_i]$ of each coordinate axis $i=1, \ldots, n$. In the absence of other information, a uniform distribution seems reasonable for the initial distribution.

b. Randomly choose N points $x^k \epsilon S$ and compute $f(x^k)$ for each. The probability distribution functions Φ_i govern these choices in that the ith component x_i^k of x^k is chosen as a random sample point from the distribution Φ_i. Thus, the Φ_i determine the allocation of trial points to various regions in S.

c. Based on the results of the trials to date, modify the Φ_i to increase the allocation of future points to regions considered likely to contain the global minimum. Go to step b.

Strategy S6 can have many realizations depending on the method of handling step c. In the version of S6 reported in this paper, step c is performed as following for each coordinate $i=1, \ldots, n$.

1. The feasible interval $[l_i, L_i]$ is split into m subintervals.

2. A "success" is defined as a value of $f(x^k)$ in the bottom 25 percent of all $f(x^k)$ values, and the ratios r_{ij} of the number of successes in subinterval j of coordinate i to the total number of points in subinterval j are computed for all i and j.

3. The modified probability for subinterval j of coordinate i is given by $p_{ij} = r_{ij} / \sum_{j=1}^{j=m} r_{ij}$ the normalized success ratio.

Several improvements on this allocation scheme are being considered for future testing.

In early tests it became apparent that performance of the various strategies fluctuated considerably, depending on the particular test problem under investigation. For example, relative to the other strategies, S2 performed spectacularly on some problems but miserably on others. On closer examination it was found that S2 did well on problems for which the global f value was significantly lower than the local minima and for which the global region of attraction was quite large; that is, on problems which were rather easy to solve. This suggests the need for a benchmark strategy to be used for assessing problem difficulty. The benchmark strategy should have as little structure as possible. We have chosen to use the pure random search method for this purpose.

STRATEGY S0 (Pure Random—Brooks [2]):

a. Set $k=1$.

b. Randomly select $x^k \epsilon S$. Evaluate $f(x^k)$.

c. Replace k with $k+1$. Go to step b. At each stage retain the best f value found to date.

This strategy may be regarded as a benchmark method since it makes no attempt to take advantage of the information gathered at previous stages. In this sense it is probably the most primitive strategy possible. We can use S0 in two ways:

1. If a strategy does not do considerably better than S0, it should be discarded.

2. If a test problem is such that S0 can solve it nearly as well as the other strategies, then the problem is not very difficult and probably is not useful for discriminating among strategies.

IV. COMPUTATIONAL EXPERIMENTS

A number of computational experiments were performed to compare the various strategies presented above. For each of the test functions employed, each strategy was run 30 times with different

random number sequences. A run was allowed to continue until the algorithm had required 1,000 evaluations of the objective function $f(x)$.

Test problems with predictable local and global solutions were constructed using the objective function

$$f(x) = -\sum_{j=1}^{j=m} c_j \exp[(x-p_j)'A_j(x-p_j)].$$

This function consists of the superposition of m modes, where mode j has depth $c_j \epsilon E^1$, position $p_j \epsilon E^n$, and shape and width determined by the $n \times n$ negative definite matrix A_j. Particular test functions were obtained by choosing the parameters c_j and p_j from a random number table. A_j was chosen to ensure that the m modes were narrow enough that they did not completely merge into one another.

Strategies S1 through S4 require an unconstrained minimizer. Since the purpose of the study is to compare global strategies, a minimizer is desired which uses the same information as is available to the other strategies—function values, but not derivatives. Powell's derivative free method was selected [4].

V. RESULTS

The computational results obtained are summarized in Tables 1 and 2. Table 1 gives characteristics of the test problems used. Table 2 lists for each problem and for each strategy the best f value obtained after 200, 500, and 1,000 function evaluations. Each value is the average of the 30 trials conducted for that problem and strategy. The percentage of the 30 trials which did not locate the global minimum after 1,000 function evaluations is also given in Table 2. It is difficult to obtain a single measure of performance for this kind of problem since we must balance speed of convergence against the chance that the global solution will be missed entirely.

TABLE 1. *Characteristics of Test Problems*

Problem	Number of variables	Number of minima	Value of global minimum
A	2	4	−9.0
B	2	10	−9.9
C	2	10	−9.3
D	2	10	−9.8
E	2	10	−13.0
F	5	5	−9.4
G	5	5	−10.1
H	5	10	−10.0
I	5	10	−8.9
J	5	20	−11.9

TABLE 2. *Test Results*

	S1	S2	S3	S4	S5	S6
	8.5	−9.0	−8.2	−8.6	−8.5	−8.7
	8.9	−9.0	−8.9	−9.0	−8.7	−9.0
	9.0	−9.0	−9.0	−9.0	−8.8	−9.0
	0.0	0.0	0.0	0.0	20.0	0.0
		−9.7	−9.0	−9.5	−9.1	−9.1
		−9.8	−9.9	−9.9	−9.7	−9.8
		−9.9	−9.9	−9.9	−9.8	−9.9
		0.0	0.0	0.0	10.0	0.0
		−8.1	−8.8	−7.8	−7.8	−7.7
		−8.2	−9.1	−8.5	−8.1	−8.0
		−8.6	−9.2	−8.7	−8.2	−8.2
		53.3	3.3	43.3	83.3	80.0
		−9.2	−7.8	−7.4	−8.8	−8.8
		−9.5	−9.4	−8.5	−9.2	−9.4
		−9.6	−9.7	−9.6	−9.2	−9.6
		30.0	6.7	33.3	73.3	33.3
		−11.8	−8.3	−9.5	−10.9	−10.2
		−12.8	−10.5	−11.2	−12.6	−12.3
		−12.9	−12.0	−13.0	−12.7	−12.8
		3.3	30.0	0.0	6.7	3.3
		−5.0	−6.4	−5.8	−0.8	−0.8
		−5.0	−8.0	−8.7	−2.9	−3.1
		−5.6	−8.5	−8.9	−7.0	−7.5
		86.7	43.3	33.3	80.0	76.7
		−7.3	−7.1	−7.5	−5.0	−4.7
		−8.8	−9.7	−9.7	−8.3	−8.2
		−9.1	−9.9	−10.1	−9.5	−9.3
		56.7	10.0	0.0	16.7	40.0
		−7.0	−6.8	−7.4	−3.7	−3.6
		−7.3	−8.7	−9.2	−6.3	−7.2
		−7.7	−9.2	−9.7	−8.2	−8.9
		93.3	56.7	20.0	60.0	50.0
		−6.3	−6.5	−6.7	−4.2	−4.2
		−7.4	−8.0	−7.8	−5.8	−5.3
		−7.6	−8.4	−8.6	−6.9	−6.1
		66.7	33.3	36.7	80.0	100.0
		−6.3	−6.7	−6.5	−3.8	−3.6
		−6.6	−7.4	−8.1	−5.3	−4.6
		−7.2	−8.8	−8.3	−7.4	−6.5
		83.3	66.7	76.7	73.3	90.0

some general conclusions:

not very challenging since S0 did nearly as well as most other

progress, but frequently stops short of the global solution—it

3. In general, S1, S3, and S4 perform about the same and better than the other strategies.

4. S5 and S6 exhibit slow initial convergence. Both frequently tend to concentrate the search effort around a good local minimum, which is not global.

5. On difficult problems even the best of these methods will frequently fail to locate the global minimum.

It is also interesting to examine the entire graph of the number of function evaulations versus the best function value obtained for each strategy. These curves are shown for test function H in Figure 1. The results for function H are representative of those obtained for the other functions and serve to emphasize conclusions 2, 3, and 4, above.

Note on this graph that S5 and S6 display a consistent decrease at an initial rate which is similar to that of the better strategies S3 and S4. However, since they start much higher on the graph, S5 and S6 never catch up. This is inherent in the methods. Given any starting point x^k, S3 and S4 immediately search to a local minimum, and thus quickly get a fairly low objective function value. Starting from the same initial point, S5 and S6 merely note the objective value and proceed to check other points, doing a global survey instead of a local minimization. Thus, in the initial stages, S5 and S6 are essentially identical to S0, pure random search. It is only after considerable information has been accumulated that these methods can concentrate their attention on promising search areas.

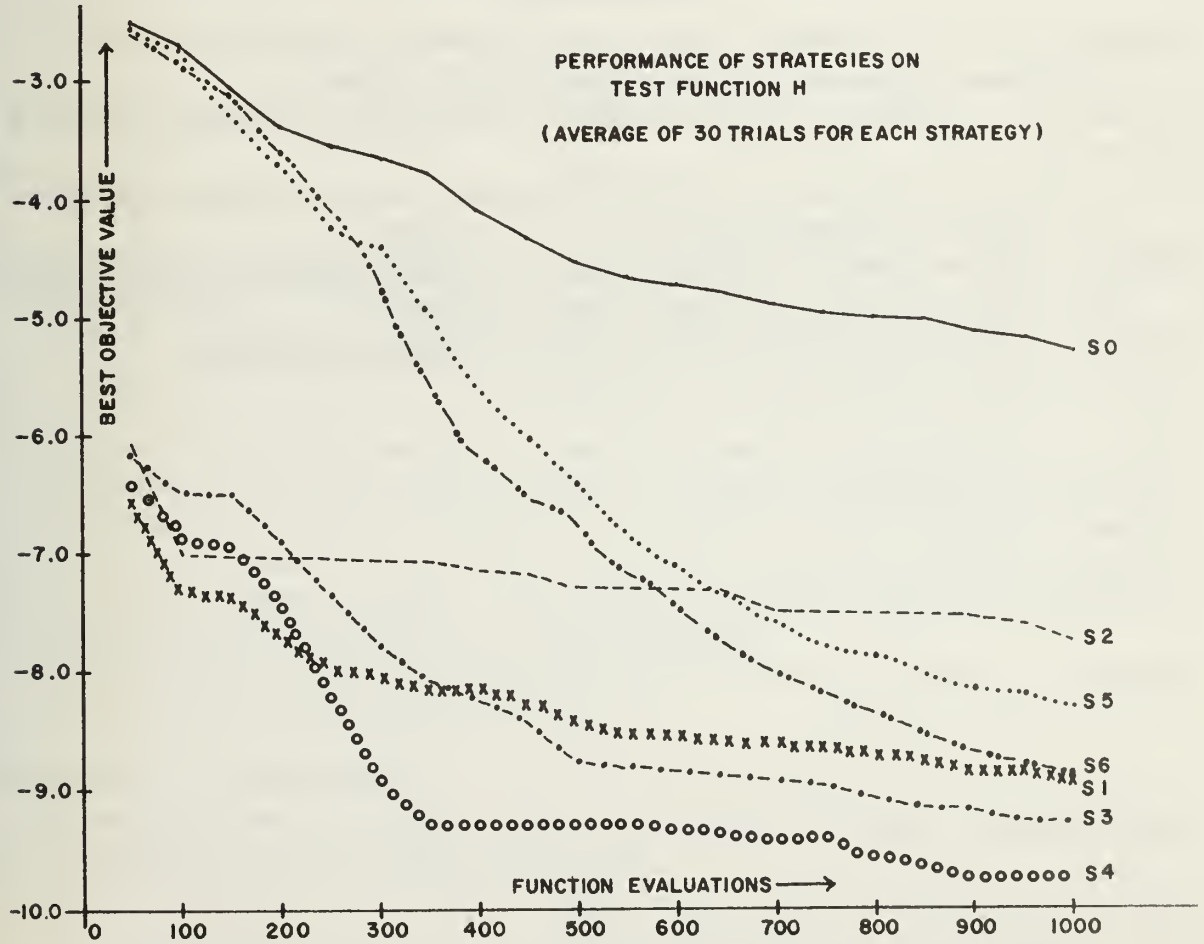

FIGURE 1. Performance of strategies on test function H (average of 30 trials for each strategy).

A comparison of S0, S1, S2, S3, and S4 is also interesting. In general, it seems that in these strategies which alternate cycles of random searching with unconstrained minimizations, the best results are obtained by the methods which do the least random searching. Thus, S0 is purely random search, and its performance is the worst. S2 requires several (perhaps many) random evaluations before each minimization to find a point better than the current best local solution, and its performance is second worst. Strategy S1 selects one random point x^k before each minimization while S3 selects one random direction d^k. Their performance is similar and almost as good as that of S4 which makes no random selections between minimizations. This strongly suggests that an improved strategy will consist of frequent minimizations coupled with an improved way of selecting starting points which are promising and which also sample the entire region.

In conclusion, it is appropriate to note that these six methods do not come near to exhausting the possible techniques for avoiding local solutions. Methods which are hybrids of these and entirely new methods should be tested. In particular, we hope to develop an algorithm which allocates unconstrained minimizations to various regions similar to the way strategy S6 allocates the individual points x^k. Such a method would combine the rapid local optimizing power of the minimization method with a global analysis of the feasible region.

REFERENCES

[1] Bocharov, N. and A. A. Feldbaum, "An Automatic Optimizer for the Search of the Smallest of Several Minima," Automation and Remote Control Vol. 23, No. 3 (1962).

[2] Brooks, S. H., "A Discussion of Random Methods for Seeking Maxima," Operations Research *6*, 244 (1958).

[3] Fiacco, A. V. and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, Inc., New York, (1968).

[4] Powell, M. J. D., "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," Computer Journal 7, 155 (1964).

[5] Zakharov, V. V., "A Random Search Method," Engineering Cybernetics, 2, 26 (1969)..

A NOTE ON MATHEMATICAL PROGRAMMING WITH FRACTIONAL OBJECTIVE FUNCTIONS

B. Mond

La Trobe University
Bundoora, Melbourne, Australia

and

B. D. Craven

University of Melbourne
Melbourne, Australia

INTRODUCTION

Consider the fractional programming problem with linear constraints. Problem 1 (P1):

(1) $$\text{Maximize } f(x)/g(x)$$

Subject to

(2) $$Ax \leqq b$$

(3) $$x \geqq 0.$$

It is assumed that the problem is regular, i.e., that the constraint set is nonempty and bounded and that f and g do not simultaneously become zero.

There has been a great deal of interest in various special cases of P1. In particular, if f and g are linear, Charnes and Cooper [1] showed that optimal solutions can be determined from optimal solutions of two associated linear programming problems. Charnes and Cooper's result was extended to the ratio of two quadratic functions by Swarup [3]. He considered P1 with f and g quadratic, and showed that an optimal solution, if it exists, can be obtained from the solutions of two associated quadratic programming problems, each with linear constraints and one quadratic constraint. Sharma [2] considered P1 with f and g polynomials. He showed that an optimal solution, if it exists, can be obtained from the solutions of two associated programming problems where the objective function is a polynomial and the constraints are all linear except for one polynomial constraint.

Here we consider a much wider class of functionals f and g, and obtain a theorem that includes as special cases the corresponding results of [1], [2], and [3].

NOTATION AND DEFINITIONS

$A \epsilon R^{m \times n}$, $x \epsilon R^n$, $b \epsilon R^m$, $t \epsilon R$, f and g are mappings from R^n into R. ϕ denotes a monotone strictly increasing function from R into R, with $\phi(t) > 0$ for $t > 0$.

For a specified function ϕ, define the functions F and G, for real positive t and $y \in R^n$, by

$$F(y, t) = f(y/t)\phi(t) \text{ and } G(y, t) = g(y/t)\phi(t). \tag{4}$$

Define

$$F(y, 0) = \lim_{t \to 0} F(y, t) \text{ and } G(y, 0) = \lim_{t \to 0} G(y, t) \tag{5}$$

whenever these limits exist. Assume that $G(0, 0) = 0$ whenever it exists.

RESULTS

Let us introduce the transformation $y = tx$, where for specified function ϕ and nonzero constant $\Delta \in R$, we require

$$G(y, t) = \Delta. \tag{6}$$

On multiplying numerator and denominator of (1) by $\phi(t)$, and using (4) and (6), we obtain the associated nonlinear programming problem. Problem 2 (P2):

$$\text{Maximize } F(y, t),$$

Subject to

$$Ay - bt \leqq 0 \tag{7}$$

$$G(y, t) = \Delta \tag{8}$$

$$y \geqq 0 \tag{9}$$

$$t \geqq 0 \tag{10}$$

LEMMA: If the point (y, t) satisfies the constraints of P2, then $t > 0$.

REMARK: This only requires proof if $G(y, 0)$ is defined, by (5). This is automatically the case if f is affine and $\phi(t) = t$.

By (8), the point (0, 0) is not feasible for P2.

PROOF: Assume that the point $(y, 0)$ is feasible for P2; then $y \neq 0$. Let x be feasible for P1. Since the constraints are linear, $x + ky$ is feasible for P1 for any positive k, contradicting the boundedness of the constraint set of P1.

THEOREM 1: If (i) $0 <$ sgn $\Delta =$ sgn $g(x^*)$ for an optimal solution x^* of P1, and (ii) (y^*, t^*) is an optimal solution of P2, then y^*/t^* is an optimal solution of P1.

PROOF: Assume that the theorem is false, i.e., assume that there exists an optimal x^* such that

$$f(x^*)/g(x^*) > f(y^*/t^*)/g(y^*/t^*). \tag{11}$$

By condition (i),

$$g(x^*) = \theta\Delta \text{ for some } \theta > 0.$$

Consider $t=\phi^{-1}(1/\theta)$, and $y=\phi^{-1}(1/\theta)x^*$. Then $\phi(t)g(x^*)=G(y,t)$ [illegible] (10), and also (7) since the constraint is linear. Thus (y,t) is a feasibl[illegible]

Now

$$f(x^*)/g(x^*)=\phi(t)f(x^*)/[\phi(t)g(x^*)]=F(y,t)/G(\quad) \tag{12}$$

Also

$$f(y^*/t^*)/g(y^*/t^*)=\phi(t^*)f(y^*/t^*)/[\phi(t^*)g(y^*/t^*)] \tag{13}$$

$$=F(y^*/t^*)/G(y^*/t^*)$$

$$=F(y^*/t^*)/\Delta.$$

Hence, for feasible (y,t), (11), (12), and (13) show that $F(y,t)>F(y$ [illegible] tion that (y^*,t^*) is optimal for P2.

If sgn $g(x^*)<0$, for x^* an optimal solution of P1, then replacing [illegible] tional is unaltered and for the new denominator we have $-g(x^*)>0$.

Thus, if P1 has a solution, it can be obtained by solving, for a [illegible] the two nonlinear programming problems. Problem 3 (P3):

$$\text{Maximize } F(y,t)$$

Subject to

$$Ay-bt\leqq 0 \tag{14}$$

$$G(y,t)=1$$

$$y\geqq 0,\ t\geqq 0$$

and Problem 4 (P4):

$$\text{Maximize } -F(y,t)$$

Subject to

$$Ay-bt\leqq 0 \tag{15}$$

$$-G(y,t)=1$$

$$y\geqq 0,\ t\geqq 0.$$

SPECIAL CASES

If f and g are linear and $\phi(t)=t$, then our theorem gives the [illegible] If f and g are quadratic functions and $\phi(t)=t^2$, we obtain the result [illegible] nomials of degree m and $\phi(t)=t^m$, we obtain the result of Sharma [2].

If f and g are homogeneous of degree k, and $\phi(t)=t^k$ then $F(y, t)=f(y)$, $G(y, t)=g(y)$, and P2 takes a simple form. An example is

$$f(x)=d^tx+(x^tCx)^{1/2},$$

where C is a positive semidefinite matrix.

REMARKS

As noted in [2] and [3], even if $G(y, t)$ is a convex function of y and t, the constraint set of P3 is not necessarily convex. Instead of P3, therefore, it is sometimes more convenient to deal with the following Problem 3′ (P3′):

$$\text{Maximize } F(y, t)$$

Subject to

$$Ay-bt\leqq 0$$

$$G(y, t)\leqq 1$$

$$y\geqq 0,\ t\geqq 0.$$

If $G(y, t)$ is a convex function of the vector variable (y, t) then the constraint set of P3′ is convex.

It should be noted that even if $g(x)$ is convex with respect to x, $G(y, t)$ need not be convex with respect to the vector variable (y, t). As an example, consider $g(x)=x^tCx-k$ where C is a positive semi-definite symmetric matrix and k is a positive scalar. Thus $g(x)$ is convex with respect to x. Taking $\phi(t)=t^2$, $G(y, t)=y^tCy-kt^2$ which is not convex.

If (y^*, t^*) is optimal for P3′, $t^*>0$, and $G(y^*, t^*)=0$, then max P1 may be ∞, since $x^*=y^*/t^*$ is feasible for P1, and $g(x^*)=0$. If $G(y^*, t^*)=\Delta_1$, where $0<\Delta_1<\infty$, then (y^*, t^*) is also optimal for P2 with $\Delta=\Delta_1$, so Theorem 1 applies. However, the optimum of P3′ can occur at $(y, t)=(0, 0)$, which does not correspond to an optimum of P1. For example, if P1 is the program (for a real variable x):

P1: Maximize $f(x)=\dfrac{-x-3}{x+1}$ subject to $x\leqq 2$ and $x\geqq 0$, then taking $\phi(t)=t$, the corresponding P3′ is:

P3′: Maximize $-y-3t$ subject to $y+t\leqq 1$, $y\geqq 0$, $t\geqq 0$, $y-2t\leqq 0$. The maximum for P3′ then occurs at $(y, t)=(0, 0)$; but the maximum for P1 occurs at $x=2$.

Similarly, instead of (P4), it might be more convenient to consider Problem 4′ (P4′)

$$\text{Maximize } -F(y, t)$$

Subject to

$$Ay-bt\leqq 0$$

$$-G(y, t)\leqq 1$$

$$y\geqq 0,\ t\geqq 0.$$

If G is concave in the vector variable (y, t), the constraint set of P4′ is convex.

REFERENCES

[1] Charnes, A. and W. W. Cooper, "Programming With Linear Fractional Functionals," Nav. Res. Log. Quart. *9*, 181–186 (1962).

[2] Sharma, I. C., "Feasible Direction Approach to Fractional Programming Problems," Opsearch *4*, 61–72 (1967).

[3] Swarup, K., "Programming with Quadratic Fractional Functionals," Opsearch *2*, 23–30 (1965).

Mathematical Models of Target Coverage and Missile Allocation

The Military Operations Research Society announces that it now has copies of its first monograph, "Mathematical Models of Target Coverage and Missile Allocation" by A. Ross Eckler and Stefan A. Burr, available for sale. The book may be purchased for $7.50 postpaid by contacting: MORS, 101 South Whiting St., Alexandria, Va. 22304.

The monograph presents a comprehensive summary of analytical models primarily used for problems in strategic defense but applicable to a wide variety of more generalized resource allocation problems. The topics discussed include models of defended point targets, circular targets, gaussian targets, generalized area targets, groups of identical targets, and nonidentical targets. Offense and defense strategies are examined and under alternative assumptions of information available to both sides. An extensive bibliography is included.

U.S. GOVERNMENT PRINTING OFFICE: 1973—541-387:1

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

SEPTEMBER 197
VOL. 20, NO. 3
NAVSO P-1278

CONTENTS

ARTICLES

		Page
Sequential Determination of Inspection Epochs for Reliability Systems with General Lifetime Distributions	S. ZACKS, W. J. FENSKE	377
An Empirical Bayes Estimator for the Scale Parameter of the Two-Parameter Weibull Distribution	G. K. BENNETT	387
Optimal Allocation of Unreliable Components for Maximizing Expected Profit Over Time	C. G. HENIN	395
A Continuous Submarine Versus Submarine Game	E. LANGFORD	405
Total Optimality of Incrementally Optimal Allocations	L. D. STONE	419
An Approach to the Allocation of Common Costs of Multi-Mission Systems	R. T. CROW	431
An Explicit General Solution in Linear Fractional Programming	A. CHARNES, W. W. COOPER	449
Using Decomposition in Integer Programming	L. SCHRAGE	469
Numerical Treatment of a Class of Semi-Infinite Programming Problems	S. A. GUSTAFSON, K. O. KORTANEK	477
Min/Max Bounds for Dynamic Network Flows	W. L. WILKINSON	505
Production-Allocation Scheduling and Capacity Expansion Using Network Flows Under Uncertainty	J. PRAWDA	517
Concave Minimization over a Convex Polyhedron	H. A. TAHA	533
Estimation of a Hidden Service Distribution of an M/G/∞ System	L. L. GEORGE, A. C. AGRAWAL	549
The Single Server Queue in Discrete Time-Numerical Analysis III	M. F. NEUTS, E. KLIMKO	557
Some Experiments in Global Optimization	J. K. HARTMAN	569
A Note on Mathematical Programming with Fractional Objective Functions	B. MOND, B. D. CRAVEN	577
News and Memoranda		583

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217